Weighted Least Squares and Adaptive Least Squares: Further Empirical Evidence

Martin Sterchi and Michael Wolf

Abstract This paper compares ordinary least squares (OLS), weighted least squares (WLS), and adaptive least squares (ALS) by means of a Monte Carlo study and an application to two empirical data sets. Overall, ALS emerges as the winner: It achieves most or even all of the efficiency gains of WLS over OLS when WLS outperforms OLS, but it only has very limited downside risk compared to OLS when OLS outperforms WLS.

1 Introduction

The linear regression model is still a cornerstone of empirical work in the social sciences. The standard textbook treatment assumes conditional homoskedasticity of the error terms. When this assumption is violated—that is, when conditional heteroskedasticity is present—standard inference is no longer valid. The current practice in such a setting is to estimate the model by ordinary least squares (OLS) and use heteroskedasticity-consistent (HC) standard errors; this approach dates back to [14].

[13] propose to 'resurrect' the previous practice of using weighted least squares (WLS), which weights the data before applying OLS. The weighting scheme is based on an estimate of the skedastic function, that is, of the function that determines the conditional variance of the error term given the values of the regressors. In practice, the model for estimating the skedastic function may be misspecified. If this is the case, using standard inference based on the weighted data will not be valid. Therefore,

M. Sterchi (✉)
University of Applied Sciences and Arts Northwestern Switzerland, Olten, Switzerland
e-mail: martin.sterchi@fhnw.ch

M. Wolf
Department of Economics, University of Zurich, Zurich, Switzerland
e-mail: michael.wolf@econ.uzh.ch

V. Kreinovich et al. (eds.), *Robustness in Econometrics*,
Studies in Computational Intelligence 692, DOI 10.1007/978-3-319-50742-2_9

[13] propose to also use HC standard errors for weighted data (as would be done for the original data) and prove asymptotic validity of the resulting inference under suitable regularity conditions.

[13] also propose adaptive least squares (ALS) where a pretest for conditional heteroskedasticity decides whether the applied researcher should use OLS (with HC standard errors) or WLS (with HC standard errors). Asymptotic validity of the resulting inference is established as well.

In addition to providing asymptotic theory, [13] examine finite-sample performance of WLS and ALS compared to OLS via Monte Carlo simulations. But these simulations are restricted to univariate regressions (that is, regressions where there is only one regressor in addition to the constant). In applied work, though, multivariate regressions are more common.

The purpose of this paper is two-fold. On the one hand, we provide extensive Monte Carlo simulations comparing WLS and ALS to OLS in multivariate regressions, covering both estimation and inference. On the other hand, we compare the results of WLS and ALS to OLS for two empirical data sets.

The remainder of the paper is organized as follows. Section 2 gives a brief description of the methodology for completeness. Section 3 examines finite-sample performance via a Monte Carlo study. Section 4 provides an application to two empirical data sets. Section 5 concludes.

2 Brief Description of the Methodology

For completeness, we give a brief description of the methodology for WLS and ALS here. More details can be found in [13].

2.1 The Model

We maintain the following set of assumptions throughout the paper.

(A1) The linear model is of the form

$$y_i = x_i'\beta + \varepsilon_i \quad (i = 1, \ldots, n), \tag{1}$$

where $x_i \in \mathbb{R}^K$ is a vector of explanatory variables (regressors), $\beta \in \mathbb{R}^K$ is a coefficient vector, and ε_i is the unobservable error term with certain properties to be specified below.

(A2) The sample $\{(y_i, x_i')\}_{i=1}^n$ is independent and identically distributed (i.i.d.).

(A3) All the regressors are predetermined in the sense that they are orthogonal to the contemporaneous error term:

$$\mathbb{E}(\varepsilon_i|x_i) = 0. \tag{2}$$

(A4) The $K \times K$ matrix $\varSigma_{xx} := \mathbb{E}(x_i x_i')$ is nonsingular (and hence finite). Furthermore, $\sum_{i=1}^n x_i x_i'$ is invertible with probability one.

(A5) The $K \times K$ matrix $\varOmega := \mathbb{E}(\varepsilon_i^2 x_i x_i')$ is nonsingular (and hence) finite.

(A6) There exists a nonrandom function $v : \mathbb{R}^K \to \mathbb{R}_+$ such that

$$\mathbb{E}(\varepsilon_i^2|x_i) = v(x_i). \tag{3}$$

Therefore, the *skedastic function* $v(\cdot)$ determines the functional form of the conditional heteroskedasticity. Note that under (A6),

$$\varOmega = \mathbb{E}\big[v(x_i) \cdot x_i x_i'\big].$$

It is useful to introduce the customary vector-matrix notations

$$y := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \varepsilon := \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad X := \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1K} \\ \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{nK} \end{bmatrix},$$

so that Eq. (1) can be written more compactly as

$$y = X\beta + \varepsilon. \tag{4}$$

Furthermore, assumptions (A2), (A3), and (A5) imply that

$$\mathrm{Var}(\varepsilon|X) = \begin{bmatrix} v(x_1) & & \\ & \ddots & \\ & & v(x_n) \end{bmatrix}.$$

2.2 Estimators: OLS, WLS, and ALS

The well-known ordinary least squares (OLS) estimator of β is given by

$$\hat{\beta}_{\mathrm{OLS}} := (X'X)^{-1}X'y.$$

Under the maintained assumptions, the OLS estimator is unbiased and consistent. This is the good news.

A more efficient estimator can be obtained by reweighting the data (y_i, x_i') and then applying OLS in the transformed model

$$\frac{y_i}{\sqrt{v(x_i)}} = \frac{x_i'}{\sqrt{v(x_i)}}\beta + \frac{\varepsilon_i}{\sqrt{v(x_i)}}. \tag{5}$$

Letting

$$V := \begin{bmatrix} v(x_1) & & \\ & \ddots & \\ & & v(x_n) \end{bmatrix},$$

the resulting estimator can be written as

$$\hat{\beta}_{\text{BLUE}} := (X'V^{-1}X)^{-1}X'V^{-1}y. \tag{6}$$

It is the best linear unbiased estimator (BLUE) and is consistent; in particular, it is more efficient than the OLS estimator. However, it is generally not a feasible estimator, since the skedastic function $v(\cdot)$ is generally unknown.

A feasible approach is to estimate the skedastic function $v(\cdot)$ from the data in some way and to then apply OLS in the model

$$\frac{y_i}{\sqrt{\hat{v}(x_i)}} = \frac{x_i'}{\sqrt{\hat{v}(x_i)}}\beta + \frac{\varepsilon_i}{\sqrt{\hat{v}(x_i)}}, \tag{7}$$

where $\hat{v}(\cdot)$ denotes the estimator of $v(\cdot)$. The resulting estimator is the weighted least squares (WLS) estimator. Letting

$$\hat{V} := \begin{bmatrix} \hat{v}(x_1) & & \\ & \ddots & \\ & & \hat{v}(x_n) \end{bmatrix},$$

the WLS estimator can be written as

$$\hat{\beta}_{\text{WLS}} := (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y.$$

It is not necessarily unbiased. If $\hat{v}(\cdot)$ is a consistent estimator of $v(\cdot)$, than WLS is asymptotically more efficient than OLS. But even if $\hat{v}(\cdot)$ is an inconsistent estimator of $v(\cdot)$, WLS can result in large efficiency gains over OLS in the presence of noticeable conditional heteroskedasticity; see Sect. 3.

The idea of adaptive least squares (ALS) is that we let the data 'decide' whether to use OLS or WLS for the estimation. Intuitively, we only want to use WLS if there is 'noticeable' conditional heteroskedasticity present in the data. Here, 'noticeable' is with respect to the model used for estimating the skedastic function in practice.

[13] suggest applying a test for conditional heteroskedasticity. Several such tests exists, the most popular ones being the tests of [2, 14]; also see [9, 10]. If the null hypothesis of conditional homoskedasticity it not rejected by such a test, use the OLS estimator; otherwise, use the WLS estimator. The resulting estimator is nothing else than the ALS estimator.

2.3 Parametric Model for Estimating the Skedastic Function

In order to estimate the skedastic function $v(\cdot)$, [13] suggest the use of the following parametric model:

$$v_\theta(x_i) := \exp\bigl(\nu + \gamma_2 \log|x_{i,2}| + \ldots + \gamma_K \log|x_{i,K}|\bigr), \tag{8}$$

with $\theta := (\nu, \gamma_2, \ldots, \gamma_K)'$, assuming that $x_{i,1} \equiv 1$ (that is, the original regression contains a constant). Otherwise, the model should be

$$v_\theta(x_i) := \exp\bigl(\nu + \gamma_1 \log|x_{i,1}| + \gamma_2 \log|x_{i,2}| + \ldots + \gamma_K \log|x_{i,K}|\bigr),$$

with $\theta := (\nu, \gamma_1, \ldots, \gamma_K)'$. Such a model is a special case of the form of multiplicative conditional heteroskedasticity previously proposed by [5] and Sect. 9.3 of [8], among others.

Assuming model (8), the test for conditional heteroskedasticity specifies

$$H_0 : \gamma_2 = \ldots = \gamma_K = 0 \quad \text{versus} \quad H_1 : \text{ at least one } \gamma_k \neq 0 \;\; (k = 2, \ldots, K).$$

To carry out the test, fix a small constant $\delta > 0$, estimate the following regression by OLS:

$$\log\bigl[\max(\delta^2, \hat{\varepsilon}_i^2)\bigr] = \nu + \gamma_2 \log|x_{i,2}| + \ldots + \gamma_K \log|x_{i,K}| + u_i, \tag{9}$$

with $\hat{\varepsilon}_i := y_i - x_i'\hat{\beta}_{\text{OLS}}$, and denote the resulting R^2-statistic by R^2.[1] Furthermore, denote by $\chi^2_{K-1,1-\alpha}$ the $1-\alpha$ quantile of the chi-squared distribution with $K-1$ degrees of freedom. Then the test rejects conditional homoskedasticity at nominal level α if $n \cdot R^2 > \chi^2_{K-1,1-\alpha}$.

Last but not least, the estimate of the skedastic function is given by

$$\hat{v}(\cdot) := v_{\hat{\theta}}(\cdot),$$

where $\hat{\theta}$ is an estimator of θ obtained by the OLS regression (9).

[1] The reason for introducing a small constant $\delta > 0$ on the left-hand side of (9) is that, because one is taking logs, one needs to avoid a residual of zero, or even very near zero. The choice $\delta = 0.1$ seems to work well in practice.

2.4 Inference: OLS, WLS, and ALS

2.4.1 Confidence Intervals

A nominal $1 - \alpha$ confidence interval for β_k based on OLS is given by

$$\hat{\beta}_{k,\text{OLS}} \pm t_{n-K,1-\alpha/2} \cdot \text{SE}_{\text{HC}}(\hat{\beta}_{k,\text{OLS}}), \tag{10}$$

where $t_{n-K,1-\alpha/2}$ denotes the $1 - \alpha/2$ quantile of the t distribution with $n - K$ degrees of freedom. Here $\text{SE}_{\text{HC}}(\cdot)$ denotes a HC standard error. Specifically [13] suggest to use the HC3 standard error introduced by [12].

A nominal $1 - \alpha$ confidence interval for β_k based on WLS is given by

$$\hat{\beta}_{k,\text{WLS}} \pm t_{n-K,1-\alpha/2} \cdot \text{SE}_{\text{HC}}(\hat{\beta}_{k,\text{WLS}}), \tag{11}$$

where again [13] suggest to use the HC3 standard error.

A nominal $1 - \alpha$ confidence interval for β_k based on ALS is given by either (10) or (11), depending on whether the ALS estimator is equal to the OLS estimator or to the WLS estimator.

2.4.2 Testing a Set of Linear Restrictions

Consider testing a set of linear restrictions on β of the form

$$H_0 : R\beta = r,$$

where $R \in \mathbb{R}^{p \times K}$ is matrix of full row rank specifying $p \leq K$ linear combinations of interest and $r \in \mathbb{R}^p$ is a vector specifying their respective values under the null.

A HC Wald statistic based on the OLS estimator is given by

$$W_{\text{HC}}(\hat{\beta}_{\text{OLS}}) := \frac{n}{p} \cdot (R\hat{\beta}_{\text{OLS}} - r)' \big[R \, \widehat{\text{Avar}}_{\text{HC}}(\hat{\beta}_{\text{OLS}}) R'\big]^{-1} (R\hat{\beta}_{\text{OLS}} - r).$$

Here $\widehat{\text{Avar}}_{\text{HC}}(\hat{\beta}_{\text{OLS}})$ denotes a HC estimator of the asymptotic variance of $\hat{\beta}_{\text{OLS}}$, that is, of the variance of the limiting multivariate normal distribution of $\hat{\beta}_{\text{OLS}}$. More specifically, if

$$\sqrt{n}(\hat{\beta}_{\text{OLS}} - \beta) \xrightarrow{d} N(0, \Sigma),$$

where the symbol $\xrightarrow{d}$ denotes convergence in distribution, then $\widehat{\text{Avar}}_{\text{HC}}(\hat{\beta}_{\text{OLS}})$ is an estimator of Σ. Related details can be found in Sect. 4 of [13]; in particular, it is again recommended to use a HC3 estimator.

A HC Wald statistic based on the WLS estimator is given by

$$W_{\mathrm{HC}}(\hat{\beta}_{\mathrm{WLS}}) := \frac{n}{p} \cdot (R\hat{\beta}_{\mathrm{WLS}} - r)' \big[R\, \widehat{\mathrm{Avar}}_{\mathrm{HC}}(\hat{\beta}_{\mathrm{WLS}}) R' \big]^{-1} (R\hat{\beta}_{\mathrm{WLS}} - r).$$

For a generic Wald statistic W, the corresponding p-value is obtained as

$$PV(W) := \mathrm{Prob}\{F \geq \tilde{W}\}, \quad \text{where } F \sim F_{p,n}.$$

Here, $F_{p,n}$ denotes the F distribution with p and n degrees of freedom.

HC inference based on the OLS estimator reports $PV(W_{\mathrm{HC}}(\hat{\beta}_{\mathrm{OLS}}))$ while HC inference based on the WLS estimator reports $PV(W_{\mathrm{HC}}(\hat{\beta}_{\mathrm{WLS}}))$. Depending on the outcome of the test for conditional heteroskedasticity, ALS inference either coincides with OLS inference (namely, if the test does not reject conditional homoskedasticity) or coincides with WLS inference (namely, if the test rejects conditional homoskedasticity).

3 Monte Carlo Evidence

3.1 Configuration

We consider the following multivariate linear regression model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i. \tag{12}$$

The regressors are first generated according to a uniform distribution between 1 and 4, denoted by $U[1, 4]$. The simulation study is then repeated with the regressors generated according to a Beta distribution with the parameters $\alpha = 2$ and $\beta = 5$, denoted by Beta(2,5). In order to guarantee a range of values comparable to the one for the uniformly distributed regressors, the Beta distributed regressors have been multiplied by five. [11] chooses a standard lognormal distribution for the regressors and points out that, as a result, HC inference becomes particularly difficult because of a few extreme observations for the regressors. Since both the standard lognormal distribution and the Beta(2,5) distribution are right-skewed, the second part of the simulation study is in the spirit of the one in [11].

The error term model in (12) is given by

$$\varepsilon_i := \sqrt{v(x_i)} z_i \tag{13}$$

where $z_i \sim N(0, 1)$ and z_i is independent of all explanatory variables x_i. Here, $v(\cdot)$ corresponds to the skedastic function and will be specified below. Alternatively, a setting with error terms following a t-distribution with five degrees of freedom (scaled

Table 1 Parametric specifications of the skedastic function

S.1	$v(x_i) = z(\gamma) \cdot \lvert x_{i,1}\rvert^{\gamma} \cdot \lvert x_{i,2}\rvert^{\gamma} \cdot \lvert x_{i,3}\rvert^{\gamma}$	with $\gamma \in \{0, 1, 2, 4\}$
S.2	$v(x_i) = z(\gamma)\big(\gamma\lvert x_{i,1}\rvert + \gamma\lvert x_{i,2}\rvert + \gamma\lvert x_{i,3}\rvert\big)$	with $\gamma \in \{1, 2, 3\}$
S.3	$v(x_i) = z(\gamma)\exp\big(\gamma\lvert x_{i,1}\rvert + \gamma\lvert x_{i,2}\rvert + \gamma\lvert x_{i,3}\rvert\big)$	with $\gamma \in \{0.5, 1\}$
S.4	$v(x_i) = z(\gamma)\big(\lvert x_{i,1}\rvert + \lvert x_{i,2}\rvert + \lvert x_{i,3}\rvert\big)^{\gamma}$	with $\gamma \in \{2, 4\}$

to have variance one) will be tested. Without loss of generality, the parameters in (12) are all set to zero, that is, $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 0, 0, 0)$.

We consider four parametric specifications of the skedastic function as shown in Table 1. For the sake of simplicity, all specifications use only one parameter γ. (For example, Specification S.1 uses a common power γ on the absolute values of $x_{i,1}$, $x_{i,2}$, and $x_{i,3}$.) It would in principle be possible to use more than one parameter in a given specification, but then the number of scenarios in our Monte Carlo study would become too large. [11] proposes the use of a scaling factor for the specifications in order to make sure that the conditional variance of ε_i is on average one, while the degree of heteroskedasticity remains the same. For that reason, all the specifications in Table 1 contain a scaling factor $z(\gamma)$. [4] suggest measuring the aforementioned degree of heteroskedasticity by the ratio of the maximal value of $v(x)$ to the minimal value of $v(x)$. Consequently, in the case of conditional homoskedasticity, the degree of heteroskedasticity is one. The full set of results is presented in Table 4; note that in specification S.2, the degree of heteroskedasticity does not depend on the value of γ.

3.2 *Estimation of the Skedastic Function*

The following parametric model is used to estimate the skedastic function:

$$v_\theta(x_i) = \exp(\upsilon + \gamma_1 \log\lvert x_{i,1}\rvert + \gamma_2 \log\lvert x_{i,2}\rvert + \gamma_3 \log\lvert x_{i,3}\rvert). \tag{14}$$

It can be reformulated as

$$v_\theta(x_i) = \exp(\upsilon) \cdot \lvert x_{i,1}\rvert^{\gamma_1} \cdot \lvert x_{i,2}\rvert^{\gamma_2} \cdot \lvert x_{i,3}\rvert^{\gamma_3}. \tag{15}$$

Formulation (15) is equivalent to specification S.1 with $\exp(\upsilon) = z(\gamma_i)$. Hence, in the case of specification S.1, we assume the correct functional form of the skedastic function when estimating it. For all other specifications mentioned in the previous section—namely S.2–S.4—model (14) is misspecified.

The parameters of model (14) will be estimated by the following OLS regression:

$$\log[\max(\delta^2, \hat{\varepsilon}_i^2)] = \upsilon + \gamma_1 \log|x_{i,1}| + \gamma_2 \log|x_{i,2}| + \gamma_3 \log|x_{i,3}| + u_i , \tag{16}$$

where the $\hat{\varepsilon}_i^2$ are the squared OLS residuals from regression (12). [13] suggest using a small constant $\delta > 0$ on the left-hand side of (16) in order to avoid taking the logarithm of squared OLS residuals near zero; as they do, we use $\delta = 0.1$.

Denote the fitted values of the regression (16) by $\hat{g}_i$. Then weights of the data for the application of WLS are simply given by $\hat{v}_i := \exp(\hat{g}_i)$, for $i = 1, \ldots, n$.

3.3 Estimation, Inference, and Performance Measures

The parameters in the regression model (12) are estimated using OLS and WLS. In addition, we include the ALS estimator. As suggested in Remark 3.1 of [13], a Breusch-Pagan test will be applied in order to determine the ALS estimator. Conditional homoskedasticity is rejected if $nR^2 > \chi^2_{3,0.9}$, where the R^2 statistic in this test is taken from the OLS regression (16). If conditional homoskedasticity is rejected, ALS coincides with WLS; otherwise ALS coincides with OLS.

To measure the performance of the different estimators, we use the empirical mean squared error (eMSE) given by

$$\text{eMSE}(\tilde{\beta}_k) := \frac{1}{B} \sum_{b=1}^{B} (\tilde{\beta}_{k,b} - \beta_k)^2, \tag{17}$$

where $\tilde{\beta}_k$ denotes a generic estimator (OLS, WLS, or ALS) of the true parameter β_k. As is well known, the population mean squared error (MSE) can be broken down into two components as follows:

$$\text{MSE}(\tilde{\beta}_k) = \text{Var}(\tilde{\beta}_k) + \text{Bias}^2(\tilde{\beta}_k). \tag{18}$$

Thus, the MSE corresponds to the sum of the variance of an estimator $\tilde{\beta}_k$ and its squared bias. While OLS is unbiased even in the case of conditional heteroskedasticity, WLS and ALS can be biased. Therefore, using the eMSE makes sure that OLS, WLS, and ALS are compared on equal footing.

We also assess the finite-sample performance of confidence intervals of the type

$$\tilde{\beta}_k \pm t_{n-4,1-\alpha/2} \cdot \text{SE}(\tilde{\beta}_k), \tag{19}$$

where SE is either the HC standard error or the maximal (Max) standard error[2] of the corresponding estimator $\tilde{\beta}_k$ and $t_{n-K,1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the t distribution with $n-K$ degrees of freedom.

First, we compute the empirical coverage probability of nominal 95% confidence intervals. Second, for OLS-Max, WLS-HC, WLS-Max, ALS-HC and ALS-Max, we compute the ratio of the average length of the confidence interval to the average length of the OLS-HC confidence interval, which thus serves as the benchmark. All the performance measures are chosen as in [13] to facilitate comparability of the results.

3.4 Results

We discuss separately the results for estimation and inference. For compactness of the exposition, we only report results for β_1. (The results for β_2 and β_3 are very similar and are available from the authors upon request.)

3.4.1 Estimation

Tables 5 and 6 in the appendix present the basic set of results when the regressors are generated according to a uniform distribution while the error terms are normally distributed. If the specification used to estimate the weights corresponds to the true specification of the skedastic function (Table 5), WLS is generally more efficient than OLS, except for the case of conditional homoskedasticity ($\gamma = 0$). For $\gamma = 0$, OLS is more efficient than WLS, which is reflected by ratios of the eMSE's (WLS/OLS) that are higher than one for all of the sample sizes. As n increases the ratios get closer to one, indicating a smaller efficiency loss of WLS compared to OLS. On the other hand, for positive values of γ, WLS is always more efficient than OLS and the efficiency gains can be dramatic for moderate and large sample sizes ($n = 50, 100$) and for noticeable conditional heteroskedasticity ($\gamma = 2, 4$). ALS offers an attractive compromise between OLS and WLS. Under conditional homoskedasticity ($\gamma = 0$), the efficiency loss compared to OLS is negligible, as all the eMSE ratios are no larger than 1.03. Under conditional heteroskedasticity, the efficiency gains over OLS are not as large as for WLS for small sample sizes ($n = 20$) but they are almost as large as for WLS for moderate sample sizes ($n = 50$) and equally as large as for WLS for large sample sizes ($n = 100$) (Tables 2 and 3).

[2]See Sect. 4.1 of [13] for a detailed description of the Max standard error. In a nutshell, the Max standard error is the maximum of the HC standard error and the 'textbook' standard error from an OLS regression, which assumes conditional homoskedasticity.

Table 2 OLS and WLS results for the CEO salaries data set. WLS/OLS denotes the ratio of the WLS-HC standard error to the OLS-HC standard error. For this data set, ALS coincides with WLS

Response variable: log($salary$)				
OLS				
Coefficient	Estimate	SE-HC	t-stat	
constant	4.504	0.290	15.54	
log($sales$)	0.163	0.039	4.15	
log($mktval$)	0.109	0.052	2.11	
ceoten	0.012	0.008	1.54	
$R^2 = 0.32$	$\bar{R}^2 = 0.31$	$s = 0.50$	$F = 26.91$	
WLS				
Coefficient	Estimate	SE-HC	t-stat	WLS/OLS
constant	4.421	0.240	18.45	0.83
log($sales$)	0.152	0.037	4.13	0.94
log($mktval$)	0.126	0.044	2.91	0.84
ceoten	0.015	0.007	2.31	0.88
$R^2 = 0.33$	$\bar{R}^2 = 0.32$	$s = 1.73$	$F = 29.04$	

Table 3 OLS and WLS results for the housing prices data set. WLS/OLS denotes the ratio of the WLS-HC standard error to the OLS-HC standard error. For this data set, ALS coincides with WLS

Response variable: log($price$)				
OLS				
Coefficient	Estimate	SE (HC)	t-stat	
constant	11.084	0.383	28.98	
log(nox)	−0.954	0.128	−7.44	
log($dist$)	−0.134	0.054	−2.48	
rooms	0.255	0.025	10.10	
stratio	−0.052	0.005	−11.26	
$R^2 = 0.58$	$\bar{R}^2 = 0.58$	$s = 0.27$	$F = 175.90$	
WLS				
Coefficient	Estimate	SE (HC)	t-stat	WLS/OLS
constant	10.195	0.272	37.43	0.71
log(nox)	−0.793	0.097	−8.17	0.76
log($dist$)	−0.127	0.035	−3.62	0.65
rooms	0.307	0.016	19.23	0.63
stratio	−0.037	0.004	−8.78	0.90
$R^2 = 0.68$	$\bar{R}^2 = 0.68$	$s = 1.33$	$F = 267.8$	

The higher the degree of heteroskedasticity, the higher the efficiency gain is of WLS over OLS. For instance, $\gamma = 4$ results in very strong conditional heteroskedasticity, as can be seen in Table 4. As a result, the ratio of the eMSE of WLS to the eMSE of OLS is below 0.05 for large sample sizes ($n = 100$). However, in the case of conditional homoskedasticity ($\gamma = 0$), OLS is more efficient than WLS, which is reflected by ratios of the eMSE's (WLS/OLS) that are higher than one for all of the sample sizes (though getting closer to one as n increases).

Figure 1 displays density plots of the three estimators of β_1 in the case of the four different parameter values of specification S.1 and for $n = 100$. The four plots visualize the potential efficiency gains of WLS and ALS over OLS as presented in Table 5 numerically. In the cases of $\gamma = 2$ and $\gamma = 4$, the density of ALS is virtually equal to the density of WLS, as there is no visible difference. It can be clearly seen how the variances of WLS and ALS get smaller relative to OLS when the degree of conditional heteroskedasticity increases.

What changes if the specification used to estimate the skedastic function does not correspond to the true specification thereof? The results for this case are presented in Table 6. First of all, the linear specification S.2 results in WLS being less efficient than OLS. Although the linear specification represents a form of conditional heteroskedasticity, it is of a different form than our parametric model used to estimate the skedastic function (that is, misspecified model). Due to the linearity of specification S.2, any choice of γ will result in the same degree of heteroskedasticity, given the sample size n. Therefore, the results of the simulation study were the same for different values of γ. Next, in specification S.3, WLS is more efficient than OLS for both choices of γ and all sample sizes. Finally, specification S.4 results in WLS being less efficient than OLS for small and moderate sample sizes ($n = 20$ and $n = 50$) and $\gamma = 2$, whereas WLS is clearly more efficient when $\gamma = 4$. Unsurprisingly, $\gamma = 4$ corresponds to a considerably higher degree of heteroskedasticity than $\gamma = 2$. Again, ALS offers an attractive compromise. It is never noticeably less efficient than OLS (that is, eMSE ratios never larger than 1.03) but is nearly as efficient ($n = 50$) or as efficient ($n = 100$) as WLS when WLS outperforms OLS.

Do the results differ if the regressors are not uniformly distributed or if the error terms are not normally distributed? In order to answer this question, the simulation study has been repeated with two different settings.

First, the regressors were chosen to follow a Beta(2,5) distribution, as specified in Sect. 3.1. As a consequence, the degree of heteroskedasticity is higher in most cases (except for specification S.3). compared to when the regressors follow a uniform distribution; see Table 4. A comparison of the two results reveals that, once again, the main factor relevant for the efficiency of WLS compared to OLS seems to be the degree of heteroskedasticity. Interestingly though, these results do not seem to apply to any degree of heteroskedasticity. Consider for example the first specification S.1. In the case of conditional homoskedasticity, the ratios of the eMSE's are similar, whereas introducing conditional heteroskedasticity ($\gamma = 1$ and $\gamma = 2$) leads to considerably stronger efficiency gains of WLS compared to OLS in the case of the Beta-distributed regressors. Unsurprisingly, the degree of heteroskedasticity for these two specifications is substantially higher in the case of Beta-distributed regressors. However, for $\gamma = 4$, WLS is more efficient in the case of uniformly distributed regressors, although the degree of heteroskedasticity is considerably lower than with

Beta-distributed regressors. The results for the other specifications (S.2–S.4) generally support the findings described in this paragraph.

Second, the basic setting has been changed by letting z_i follow a t-distribution with five degrees of freedom (scaled to have variance one). For small and moderate sample sizes ($n = 20, 50$), the efficiency gains of WLS over OLS are more pronounced compared to normally distributed z_i, whereas the efficiency gains are similar for $n = 100$.

As before, ALS offers an attractive compromise: (i) it is never noticeably less efficient than OLS and (ii) it enjoys most ($n = 50$) or practically all ($n = 100$) of the efficiency gains of WLS in case WLS outperforms OLS.

Remark 1 (Graphical Comparison) We find it useful to 'condense' the information on the ratios of the eMSE's contained in Tables 5, 6, 7, 8, 9 and 10 into a single Fig. 2. For each sample size ($n = 20, 50, 100$) and each method (WLS and ALS) there are 27 eMSE ratios compared to OLS. Here the number 27, corresponds to all combinations of specification of the skedastic function, corresponding parameter, distribution of the regressors, and distribution of the error term. For each sample size ($n = 20, 50, 100$), two boxplots are juxtaposed: one for the 27 eMSE ratios of WLS and one for the 27 eMSE ratios of ALS. In each case, a dashed horizontal line indicates the value of 1.0 (that is, same efficiency as OLS).

It can be seen that for each sample size, ALS has smaller risk of efficiency loss (with respect to OLS) than WLS: the numbers above the horizontal 1.0-line do not extend as far up. On the other hand, ALS also has a smaller chance of efficiency gain (with respect to OLS) than WLS: the numbers below the horizontal 1.0-line do not extend as far down. But the corresponding differences diminish with the sample size: There is a marked difference for $n = 20$, a moderate difference for $n = 50$, and practically no difference for $n = 100$.

Therefore, it can also be seen graphically that ALS offers an attractive compromise: (i) it is never noticeably less efficient than OLS and (ii) it enjoys most ($n = 50$) or practically all ($n = 100$) of the efficiency gains of WLS in case WLS outperforms OLS. □

3.4.2 Inference

As described in Sect. 3.3, we use two performance measures to evaluate confidence intervals: the empirical coverage probability of a nominal 95% confidence interval and the ratio of the average length of a confidence interval to the average length of the OLS-HC confidence interval.[3]

The results for the basic setting, in which the regressors are uniformly distributed and the error terms are normally distributed, are presented in Tables 11 and 12.

[3]The second performance measure does not depend on the nominal confidence level, since by definition (19), it is equivalent to the ratio of the average standard error of a given method to the average OLS-HC standard error.

In general, confidence intervals based on WLS-HC standard errors tend to undercover for small and moderate sample sizes ($n = 20, 50$). The empirical coverage probabilities for the OLS-HC confidence intervals, on the other hand, are generally satisfactory. Based on the theory, we would expect that all the HC confidence intervals tend to undercover in small samples due to the bias and increased variance of HC standard error estimates. Yet, the results here indicate that the HC confidence intervals for the WLS estimator are more prone to liberal inference. [6, p. 137] points out that the large-sample approximations for WLS are often unsatisfactory because WLS requires the estimation of more parameters (the parameters of the skedastic function) than OLS. Increasing the sample size improves the adequacy of the WLS-HC confidence intervals and the empirical coverage probabilities are always above 94% for $n = 100$. ALS-HC confidence intervals exhibit better coverage than WLS-HC confidence intervals: Already for $n = 50$, the empirical coverage probabilities are always over 94%.

When the degree of heteroskedasticity is high, then the average length of WLS-HC confidence intervals can be substantially shorter than the average length of OLS-HC confidence intervals. For instance, for specification S.1 with $\gamma = 4$ and $n = 100$, the average length of the WLS-HC confidence interval amounts to only 18% of the average length of the OLS-HC confidence interval, while the empirical coverage probability is more than satisfactory (95.8%). It is important to note that on average short confidence intervals are only desirable if, at the same time, the empirical coverage probability is satisfactory. These findings have important implications for empirical research. It is crucial to only apply WLS in combination with HC standard errors when the sample size is large enough, that is, $n \geq 100$. For smaller sample sizes, the results of the simulation study have shown that the empirical coverage probabilities can be too low. On the other hand, the ALS-HC confidence interval appears trustworthy for moderate sample sizes already, that is, for $n \geq 50$. Furthermore, the efficiency gains of the ALS-HC confidence interval over the OLS-HC (in terms of average length) are generally also substantial in the presence of noticeable conditional heteroskedasticity. For instance, for specification S.1 with $\gamma = 4$ and $n = 100$, the average length of the ALS-HC confidence interval also amounts to only 18% of the average length of the OLS-HC confidence interval, while the empirical coverage probability is more than satisfactory (95.8%).

As before, we want to analyze what happens when the regressors follow a Beta distribution as specified in Sect. 3.1, instead of a uniform distribution. As can be seen in Tables 13 and 14, for most of the specifications, the WLS-HC confidence intervals do not have a satisfactory empirical coverage probability, especially for small sample sizes. In the case of S.1 with $\gamma = 2$ or $\gamma = 4$, however, the empirical coverage probability is surprisingly high even for small sample sizes. [3] note that in the case of severe heteroskedasticity, the HC standard errors might be upward biased. In fact, the degree of heteroskedasticity is quite extreme for these two specifications and it is much higher than in the case of uniformly distributed regressors; see Table 4.

In contrast to the WLS-HC confidence intervals, the ALS-HC confidence intervals exhibit satisfactory coverage for moderate and large sample sizes ($n = 50, 100$) with all empirical coverage probabilities exceeding 94%.

The main result shown in [3] is that the bias of HC standard errors not only depends on the sample size, but also on whether or not a sample contains high leverage points. In empirical work, an observation is usually considered as a high leverage point if its diagonal element of the hat matrix is larger than $2p/n$, where p is the rank of the design matrix X.[4] A comparison of the diagonal elements of the hat matrix for both distributional assumptions of the regressors reveals that the samples created by Beta-distributed regressors generally contain more high leverage points. For instance, when $n = 100$, the sample with Beta distributed regressors contains six high leverage points, while the sample with uniformly distributed regressors only contains two high leverage points. Interestingly, for $n = 100$, the empirical coverage probability, for both OLS-HC and WLS-HC, is always larger for uniformly distributed regressors, that is, samples with fewer high leverage points, except for S.1 with $\gamma = 2, 4$ (which was discussed above).

Remark 2 (Maximal Standard Errors) The problem of undercoverage for small and moderate sample sizes (=20, 50) can be mitigated by using maximal standard errors, that is, by the use of WLS-Max and ALS-Max. Using maximal standard errors is proposed in Sect. 8.1 of [1], for example. However, these intervals can overcover by a lot for large sample sizes ($n = 100$), exhibiting empirical coverage probabilities sometimes near 100%. (This is also true for OLS-Max, although to a lesser extent.) Therefore, using maximal standard errors to mitigate undercoverage for small and moderate sample sizes seems a rather crude approach. A more promising approach, not leading to sizeable overcoverage for large sample sizes, would be the use of bootstrap methods. This topic is currently under study. □

Remark 3 (Graphical Comparison) We find it useful to 'condense' the information on the ratios of the average lengths of confidence intervals contained in Tables 11, 12, 13, 14, 15 and 16 into a single Fig. 3. We only do this for the sample size $n = 100$ to ensure a fair comparison. Comparisons for $n = 20, 50$ would not be really fair to OLS, given that WLS confidence intervals tend to undercover for $n = 20, 50$ and that ALS confidence intervals tend to undercover for $n = 20$.

It can be seen that both WLS and ALS are always weakly more efficient than OLS in the sense that none of the average-length ratios are above 1.0. It can also be seen that, for all practical purposes, ALS is as efficient as OLS. □

4 Empirical Applications

This section examines the application of OLS, WLS, and ALS to two empirical data sets. As will be seen the use of WLS and ALS can lead to much smaller standard

[4] It can be shown [7, e.g.] that p/n corresponds to the average element of the hat matrix.

errors (and thus much shorter confidence intervals) in the presence of noticeable conditional heteroskedasticity.

The two data sets are taken from [15].[5] In the first example, we model CEO salaries while in the second example, we model housing prices.

4.1 CEO Salaries

This cross-sectional data set from 1990 contains the salaries of 177 CEOs as well as further variables describing attributes of the CEOs and the corresponding companies. The model considered in this section tries to explain the log of the CEO salaries.[6] The variables (one response and three explanatory) used in the regression model under consideration are as follows:

$\log(salary)$:	log of CEO's salary (in US$1,000)
$\log(sales)$:	log of firm sales (in million US$)
$\log(mktval)$:	log of market value (in million US$)
$ceoten$:	years as the CEO of the company

The sample size is $n = 177$ and the number of regressors (including the constant) is $K = 4$. Based on the results of the Monte Carlo study in Sect. 3, the sample size is large enough so that WLS and ALS inference can both be trusted.

The model is specified as in [15, p. 213] and is first estimated using OLS. The results are shown in the upper part of Table 2. The estimated coefficients are all positive, which intuitively makes sense. Examining the t-statistics (based on HC standard errors) shows that all estimated coefficients are significant at the 5% level except for the estimated coefficient on *ceoten*, which is insignificant.

The lower part of Table 2 shows the WLS results. The WLS estimates do not substantially differ from the OLS estimates. However, the HC standard errors are always smaller for WLS compared to OLS and generally noticeably so, with the ratios ranging from 0.93 to 0.84. In particular, now all estimated coefficients are individually significant at the 5% level, including the estimated coefficient on *ceoten*.

To determine the nature of ALS, we run a Breusch-Pagan test as described in Sect. 2.3.[7] The critical value of the test is $\chi^2_{3,0.90} = 6.25$ and the value of the test statistic is 8.25. Hence, the test detects conditional heteroskedasticity and ALS coincides with WLS.

---

[5]The two data sets are available under the names CEOSAL2 and HPRICE2, respectively at http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html.

[6]The log always corresponds to the natural logarithm.

[7]This regression results in taking the log of $\log(sales)$ and $\log(mktval)$ on the right-hand side; taking absolute values is not necessary, since $\log(sales)$ and $\log(mktval)$ are always positive. Furthermore, some observations have a value of zero for $ceoten$; we replace those values by 0.01 before taking logs.

4.2 Housing Prices

This cross-sectional data set from 1970 contains 506 observations from communities in the Boston area. The aim is to explain the median housing price in a community by means of the level of air pollution, the average number of rooms per house and other community characteristics. The variables (one response and four explanatory) used in the regression model under consideration are as follows:

$\log(price)$:	log of median housing price (in US$)
$\log(nox)$:	log of nitrogen oxide in the air (in parts per million)
$\log(dist)$:	log of weighted distance from 5 employment centers (in miles)
$rooms$:	average number of rooms per house
$stratio$:	average student-teacher ratio

The sample size is $n = 506$ and the number of regressors (including the constant) is $K = 5$. Based on the results of the Monte Carlo study in Sect. 3, the sample size is large enough so that WLS and ALS inference can both be trusted.

The model follows an example in [15, p. 132]. The results from the OLS estimation are reported in the upper part of Table 3. All the estimated coefficients have the expected sign and are significant at the 1% level.

The lower part of Table 3 shows the WLS results. The WLS estimates do not substantially differ from the OLS estimates. However, the HC standard errors are always smaller for WLS compared to OLS and generally noticeably so, with the ratios ranging from 0.90 to 0.63. As for OLS, all estimated coefficients are significant at the 1% level. But the corresponding confidence intervals based on WLS are shorter compared to OLS due to the smaller standard errors, which results in more informative inference. For example, a 95% confidence interval for the coefficient on *rooms* is given by [0.276, 0.338] based on WLS and by [0.258, 0.356] based on OLS. Needless to say, the smaller standard errors for WLS compared to OLS would also result in more powerful hypothesis tests concerning the various regression coefficients.

To determine the nature of ALS, we run a Breusch-Pagan test as described in Sect. 2.3. The critical value of the test is $\chi^2_{4,0.90} = 7.78$ and the value of the test statistic is 92.08. Hence, the test detects conditional heteroskedasticity and ALS coincides with WLS.

5 Conclusion

The linear regression model remains a cornerstone of applied research in the social sciences. Many real-life data sets exhibit conditional heteroskedasticity which makes text-book inference based on ordinary least squares (OLS) invalid. The current practice in analyzing such data sets—going back to [14]—is to use OLS in conjunction with heteroskedasticity consistent (HC) standard errors.

In a recent paper, [13] suggest to return to the previous practice of using weighted least squares (WLS), also in conjunction with HC standard errors. Doing so ensures validity of the resulting inference even if the model for estimating the skedastic function is misspecified. In addition, they make the new proposal of adaptive least squares (ALS), where it is 'decided' from the data whether the applied researcher should use either OLS or WLS, in conjunction with HC standard errors.

This paper makes two contributions. On the one hand, we have compared finite-sample performance of OLS, WLS, and ALS for multivariate regressions via a Monte Carlo study. On the other hand, we have compared OLS, WLS, and ALS when applied to two empirical data sets.[8]

The results of the Monte Carlo study point towards ALS as the overall winner. When WLS outperforms OLS, then ALS achieves most (for moderate sample sizes) or even all (for large sample sizes) of the gains of WLS; and these gains can be dramatic. When OLS outperforms WLS, then it also outperforms ALS but by a much smaller margin. Consequently, when comparing ALS to OLS, there is large upside potential and only very limited downside risk.

The application to two empirical data sets have shown that WLS and ALS can achieve large efficiency gains over OLS in the presence of noticeable conditional heteroskedasticity. Namely, smaller standard errors result in shorter (and thus more informative) confidence intervals and in more powerful hypothesis tests.

A Figures and Tables

[8] [13] only use univariate regressions in their Monte Carlo study and do not provide any applications to empirical data sets.

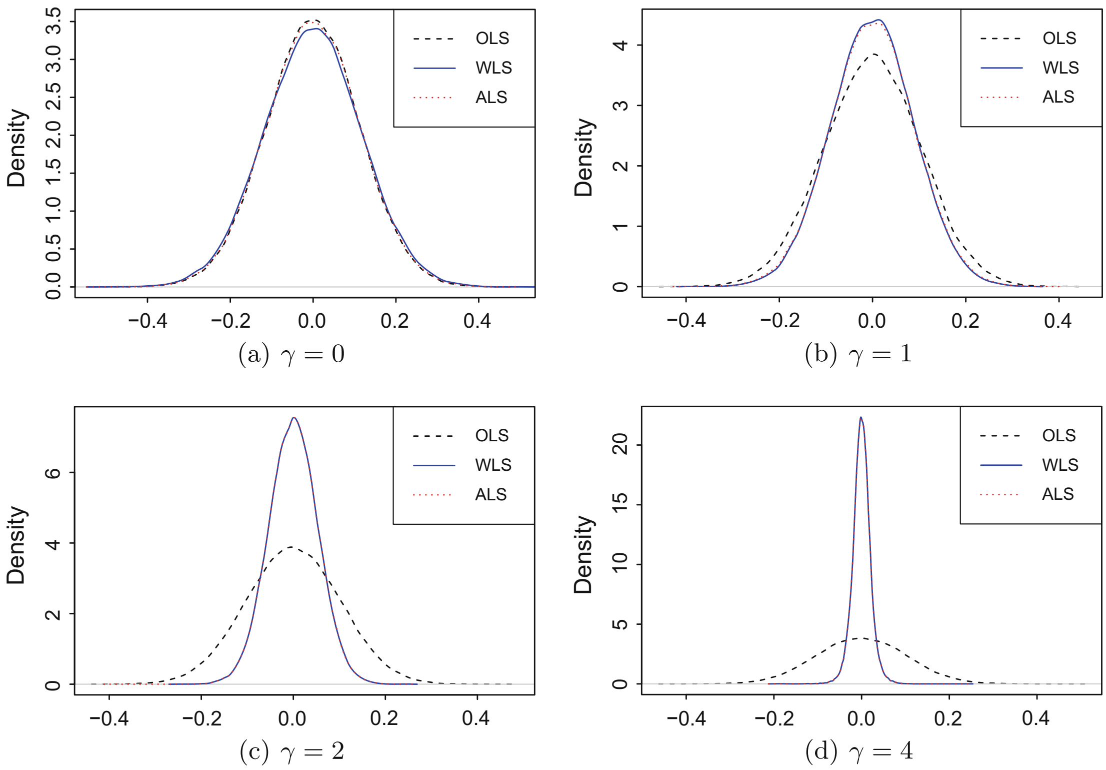

Fig. 1 Density plots for the estimators of β_1 for Specification S.1 and its four parameter values. The sample size is 100, the regressors are $U[1, 4]$-distributed and the error terms follow a standard normal distribution

Table 4 Degree of heteroskedasticity for the different specifications of the scedastic function. The degree of heteroskedasticity is measured as $\max(v(x))/\min(v(x))$

	S.1		S.2		S.3		S.4	
	Uniform	Beta	Uniform	Beta	Uniform	Beta	Uniform	Beta
	$\gamma = 0$		$\gamma = 1$		$\gamma = 0.5$		$\gamma = 2$	
$n = 20$	1.0	1.0	1.9	3.0	14.4	10.0	3.8	8.7
$n = 50$	1.0	1.0	2.0	5.3	15.2	24.0	4.1	28.0
$n = 100$	1.0	1.0	2.8	6.4	34.0	25.2	7.9	41.1
	$\gamma = 1$		$\gamma = 2$		$\gamma = 1$		$\gamma = 4$	
$n = 20$	9.7	174.1	1.9	3.0	206.2	99.5	14.3	76.0
$n = 50$	10.4	439.3	2.0	5.3	231.8	576.9	16.5	781.9
$n = 100$	24.3	682.5	2.8	6.4	1,157.5	633.8	62.5	1,689.3
	$\gamma = 2$		$\gamma = 3$					
$n = 20$	93.3	30,323.5	1.9	3.0				
$n = 50$	108.3	193,011.0	2.0	5.3				
$n = 100$	590.5	465,764.5	2.8	6.4				
	$\gamma = 4$							
$n = 20$	8,699.6	0.92×10^9						
$n = 50$	11,737.4	37×10^9						
$n = 100$	348,646.3	217×10^9						

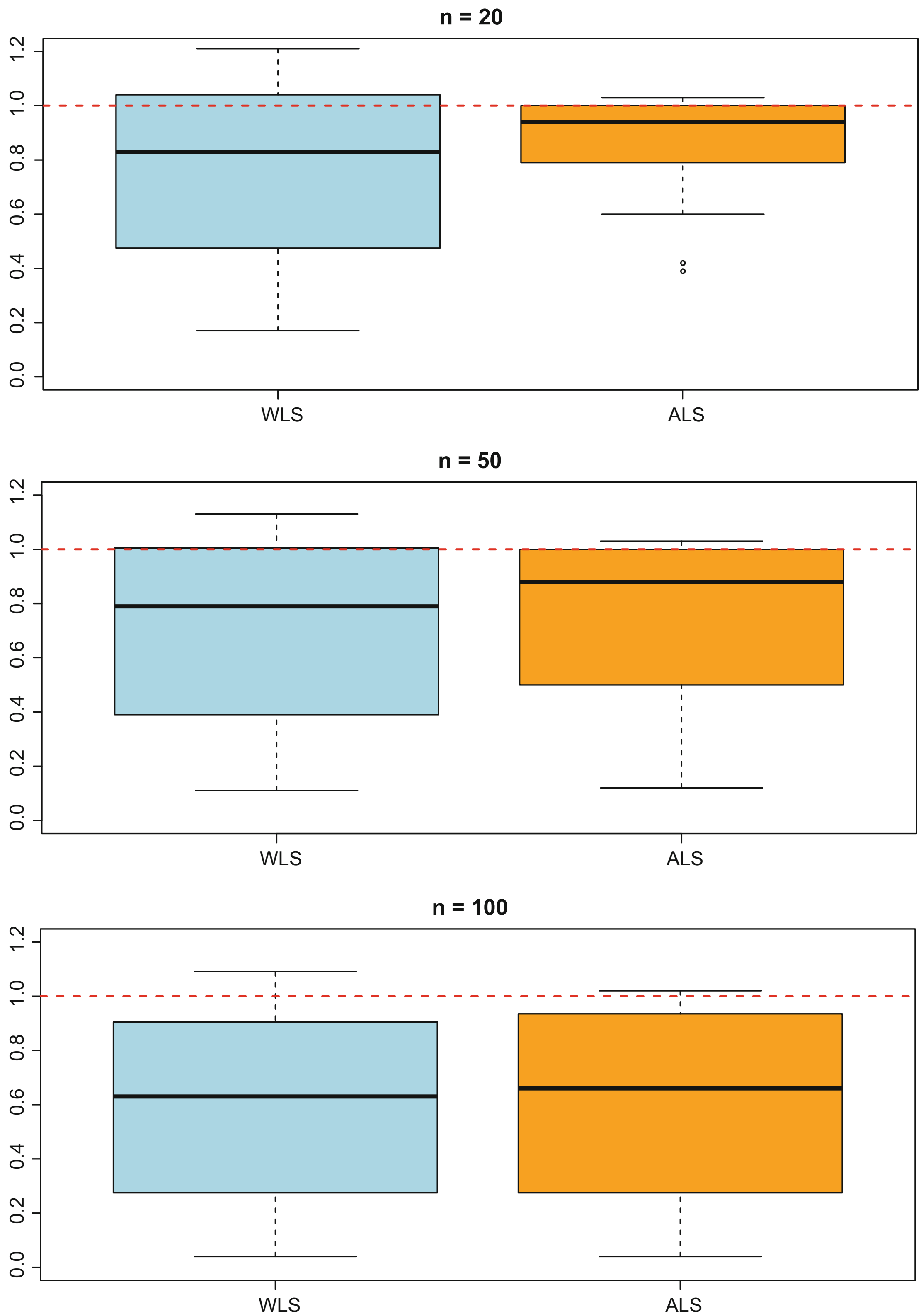

Fig. 2 Boxplots of the ratios of the eMSE of WLS (*left*) and ALS (*right*) to the eMSE of OLS. For a given sample size $n = 20, 50, 100$, the boxplots are over all 27 combinations of specification of the skedastic function, parameter value, distribution of the regressors, and distribution of the error terms

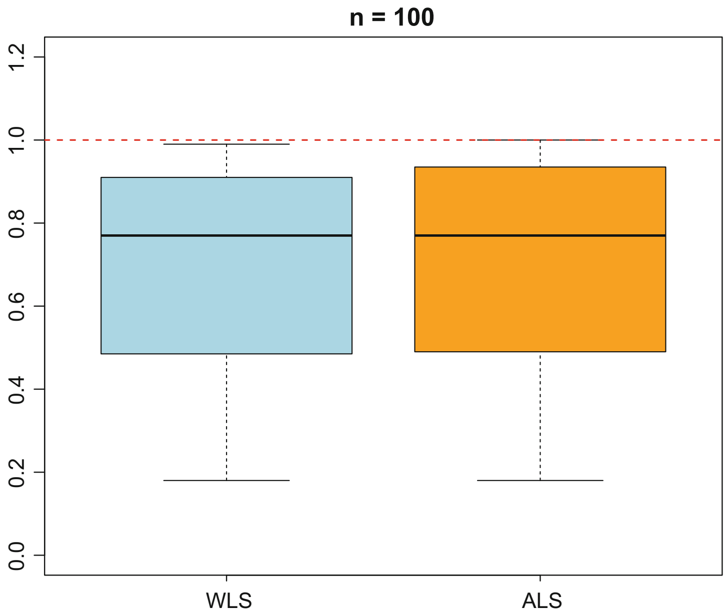

Fig. 3 Boxplots of the ratios of the average length of WLS confidence intervals for β_1 (*left*) and ALS confidence intervals for β_1 (*right*) to the average length of OLS confidence intervals for β_1. For the given sample size $n = 100$, the boxplots are over all 27 combinations of specification of the skedastic function, parameter value, distribution of the regressors, and distribution of the error terms

Table 5 Empirical mean squared errors (eMSEs) of estimators of β_1 in the case of Specification S.1. The numbers in parentheses express the ratios of the eMSE of a given estimator to the eMSE of the OLS estimator. The regressors are $U[1, 4]$-distributed and the error terms follow a standard normal distribution.

	OLS	WLS	ALS
S.1 ($\gamma = 0$)			
$n = 20$	0.064	0.077 (1.19)	0.066 (1.03)
$n = 50$	0.029	0.032 (1.13)	0.029 (1.03)
$n = 100$	0.013	0.014 (1.08)	0.013 (1.02)
S.1 ($\gamma = 1$)			
$n = 20$	0.071	0.065 (0.92)	0.070 (0.98)
$n = 50$	0.026	0.022 (0.85)	0.025 (0.93)
$n = 100$	0.011	0.008 (0.72)	0.008 (0.73)
S.1 ($\gamma = 2$)			
$n = 20$	0.084	0.042 (0.50)	0.062 (0.73)
$n = 50$	0.028	0.012 (0.42)	0.014 (0.49)
$n = 100$	0.010	0.003 (0.27)	0.003 (0.27)
S.1 ($\gamma = 4$)			
$n = 20$	0.097	0.019 (0.20)	0.041 (0.42)
$n = 50$	0.034	0.004 (0.10)	0.004 (0.12)
$n = 100$	0.010	0.000 (0.04)	0.000 (0.04)

Table 6 Empirical mean squared errors (eMSEs) of estimators of β_1 in the case of Specifications S.2–S.4. The numbers in parentheses express the ratios of the eMSE of a given estimator to the eMSE of the OLS estimator. The regressors are $U[1, 4]$-distributed and the error terms follow a standard normal distribution

	OLS	WLS	ALS
S.2 ($\gamma > 0$)			
$n = 20$	0.066	0.077 (1.17)	0.068 (1.03)
$n = 50$	0.028	0.030 (1.10)	0.028 (1.03)
$n = 100$	0.012	0.013 (1.04)	0.012 (1.02)
S.3 ($\gamma = 0.5$)			
$n = 20$	0.077	0.064 (0.83)	0.073 (0.94)
$n = 50$	0.028	0.022 (0.79)	0.024 (0.88)
$n = 100$	0.011	0.007 (0.65)	0.008 (0.67)
S.3 ($\gamma = 1$)			
$n = 20$	0.092	0.036 (0.39)	0.058 (0.63)
$n = 50$	0.030	0.010 (0.33)	0.012 (0.39)
$n = 100$	0.011	0.002 (0.20)	0.002 (0.20)
S.4 ($\gamma = 2$)			
$n = 20$	0.069	0.074 (1.08)	0.070 (1.02)
$n = 50$	0.027	0.028 (1.03)	0.027 (1.01)
$n = 100$	0.012	0.011 (0.92)	0.011 (0.93)
S.4 ($\gamma = 4$)			
$n = 20$	0.076	0.063 (0.83)	0.072 (0.94)
$n = 50$	0.027	0.021 (0.79)	0.024 (0.88)
$n = 100$	0.011	0.007 (0.61)	0.007 (0.62)

Table 7 Empirical mean squared errors (eMSEs) of estimators of β_1 in the case of Specification S.1. The numbers in parentheses express the ratios of the eMSE of a given estimator to the eMSE of the OLS estimator. The regressors are Beta(2,5)-distributed and the error terms follow a standard normal distribution

	OLS	WLS	ALS
S.1 ($\gamma = 0$)			
$n = 20$	0.142	0.172 (1.21)	0.147 (1.03)
$n = 50$	0.032	0.037 (1.13)	0.033 (1.03)
$n = 100$	0.013	0.014 (1.09)	0.013 (1.02)
S.1 ($\gamma = 1$)			
$n = 20$	0.122	0.081 (0.66)	0.106 (0.87)
$n = 50$	0.034	0.016 (0.46)	0.020 (0.58)
$n = 100$	0.016	0.006 (0.36)	0.006 (0.36)
S.1 ($\gamma = 2$)			
$n = 20$	0.129	0.049 (0.38)	0.095 (0.74)
$n = 50$	0.033	0.006 (0.18)	0.010 (0.31)
$n = 100$	0.017	0.002 (0.13)	0.002 (0.13)
S.1 ($\gamma = 4$)			
$n = 20$	0.136	0.038 (0.28)	0.115 (0.84)
$n = 50$	0.025	0.003 (0.13)	0.013 (0.52)
$n = 100$	0.014	0.003 (0.18)	0.003 (0.19)

Table 8 Empirical mean squared errors (eMSEs) of estimators of β_1 in the case of Specifications S.2–S.4. The numbers in parentheses express the ratios of the eMSE of a given estimator to the eMSE of the OLS estimator. The regressors are Beta(2,5)-distributed and the error terms follow a standard normal distribution

	OLS	WLS	ALS
S.2 ($\gamma > 0$)			
$n = 20$	0.131	0.152 (1.16)	0.134 (1.02)
$n = 50$	0.033	0.035 (1.04)	0.034 (1.01)
$n = 100$	0.014	0.014 (0.97)	0.014 (0.99)
S.3 ($\gamma = 0.5$)			
$n = 20$	0.123	0.121 (0.99)	0.122 (0.99)
$n = 50$	0.035	0.029 (0.81)	0.032 (0.91)
$n = 100$	0.018	0.013 (0.70)	0.014 (0.72)
S.3 ($\gamma = 1$)			
$n = 20$	0.111	0.070 (0.63)	0.098 (0.88)
$n = 50$	0.036	0.013 (0.37)	0.018 (0.50)
$n = 100$	0.025	0.007 (0.28)	0.007 (0.28)
S.4 ($\gamma = 2$)			
$n = 20$	0.123	0.124 (1.01)	0.123 (1.00)
$n = 50$	0.035	0.029 (0.82)	0.032 (0.92)
$n = 100$	0.016	0.012 (0.72)	0.012 (0.74)
S.4 ($\gamma = 4$)			
$n = 20$	0.115	0.079 (0.69)	0.103 (0.89)
$n = 50$	0.037	0.016 (0.44)	0.020 (0.54)
$n = 100$	0.021	0.007 (0.33)	0.007 (0.33)

Table 9 Empirical mean squared errors (eMSEs) of estimators of β_1 in the case of Specification S.1. The numbers in parentheses express the ratios of the eMSE of a given estimator to the eMSE of the OLS estimator. The regressors are $U[1, 4]$-distributed but the error terms follow a t-distribution with five degrees of freedom

	OLS	WLS	ALS
S.1 ($\gamma = 0$)			
$n = 20$	0.064	0.070 (1.10)	0.064 (1.01)
$n = 50$	0.028	0.030 (1.08)	0.029 (1.01)
$n = 100$	0.013	0.013 (1.04)	0.013 (1.01)
S.1 ($\gamma = 1$)			
$n = 20$	0.071	0.060 (0.84)	0.067 (0.94)
$n = 50$	0.026	0.022 (0.82)	0.024 (0.91)
$n = 100$	0.011	0.008 (0.70)	0.008 (0.72)
S.1 ($\gamma = 2$)			
$n = 20$	0.084	0.038 (0.45)	0.058 (0.70)
$n = 50$	0.028	0.011 (0.41)	0.014 (0.50)
$n = 100$	0.011	0.003 (0.28)	0.003 (0.28)
S.1 ($\gamma = 4$)			
$n = 20$	0.096	0.016 (0.17)	0.038 (0.39)
$n = 50$	0.034	0.004 (0.11)	0.004 (0.13)
$n = 100$	0.011	0.001 (0.05)	0.001 (0.05)

Table 10 Empirical mean squared errors (eMSEs) of estimators of β_1 in the case of Specification S.2–S.4. The numbers in parentheses express the ratios of the eMSE of a given estimator to the eMSE of the OLS estimator. The regressors are $U[1, 4]$-distributed but the error terms follow a t-distribution with five degrees of freedom

	OLS	WLS	ALS
S.2 ($\gamma > 0$)			
$n = 20$	0.065	0.070 (1.07)	0.066 (1.00)
$n = 50$	0.027	0.029 (1.05)	0.028 (1.01)
$n = 100$	0.012	0.012 (1.00)	0.012 (1.00)
S.3 ($\gamma = 0.5$)			
$n = 20$	0.077	0.058 (0.76)	0.069 (0.90)
$n = 50$	0.027	0.021 (0.76)	0.024 (0.87)
$n = 100$	0.012	0.007 (0.63)	0.008 (0.66)
S.3 ($\gamma = 1$)			
$n = 20$	0.091	0.031 (0.35)	0.055 (0.60)
$n = 50$	0.030	0.010 (0.32)	0.012 (0.40)
$n = 100$	0.011	0.002 (0.21)	0.002 (0.21)
S.4 ($\gamma = 2$)			
$n = 20$	0.068	0.068 (1.00)	0.067 (0.99)
$n = 50$	0.027	0.026 (0.98)	0.027 (0.99)
$n = 100$	0.012	0.011 (0.89)	0.011 (0.94)
S.4 ($\gamma = 4$)			
$n = 20$	0.076	0.058 (0.76)	0.068 (0.90)
$n = 50$	0.027	0.020 (0.76)	0.023 (0.87)
$n = 100$	0.012	0.007 (0.60)	0.007 (0.62)

Table 11 Empirical coverage probabilities of nominal 95% confidence intervals for β_1 in the case of Specification S.1 (in percent). The numbers in parentheses express the ratios of the average length of a given confidence interval to the average length of OLS-HC. The regressors are $U[1, 4]$-distributed and the error terms follow a standard normal distribution

	OLS-HC	OLS-Max	WLS-HC	WLS-Max	ALS-HC	ALS-Max
S.1 ($\gamma = 0$)						
$n = 20$	96.4	97.1 (1.02)	92.7 (0.91)	93.6 (0.93)	95.4 (0.98)	96.1 (1.00)
$n = 50$	95.5	96.2 (1.02)	93.3 (0.97)	94.1 (0.99)	94.9 (0.99)	95.7 (1.01)
$n = 100$	95.4	95.9 (1.02)	94.1 (0.99)	94.7 (1.01)	95.1 (1.00)	95.6 (1.01)
S.1 ($\gamma = 1$)						
$n = 20$	96.6	97.1 (1.01)	93.0 (0.81)	93.9 (0.82)	95.0 (0.91)	95.6 (0.93)
$n = 50$	95.7	96.7 (1.04)	93.9 (0.85)	94.7 (0.88)	94.2 (0.91)	95.1 (0.93)
$n = 100$	95.5	96.7 (1.06)	94.3 (0.81)	95.2 (0.84)	94.1 (0.82)	95.1 (0.85)
S.1 ($\gamma = 2$)						
$n = 20$	96.3	96.6 (1.00)	92.9 (0.58)	93.9 (0.59)	94.4 (0.70)	94.4 (0.71)
$n = 50$	95.4	96.3 (1.03)	94.1 (0.60)	95.1 (0.62)	94.3 (0.62)	94.8 (0.64)
$n = 100$	95.4	97.2 (1.09)	94.3 (0.50)	96.7 (0.56)	94.3 (0.50)	96.7 (0.56)
S.1 ($\gamma = 4$)						
$n = 20$	96.1	96.2 (1.00)	94.2 (0.31)	94.9 (0.32)	94.2 (0.40)	94.8 (0.41)
$n = 50$	94.8	95.2 (1.01)	94.6 (0.27)	97.1 (0.31)	94.5 (0.27)	97.0 (0.31)
$n = 100$	95.7	97.6 (1.11)	95.8 (0.18)	99.9 (0.32)	95.8 (0.18)	99.9 (0.32)