Heteroskedasticity-robust tests in linear regression: A review and evaluation of small-sample corrections

James E. Pustejovsky and Gleb Furman

The University of Texas at Austin

Author Note

James E. Pustejovsky, Department of Educational Psychology, University of Texas at Austin; Gleb Furman, Department of Educational Psychology, University of Texas at Austin.

An earlier version of this paper was presented at the 2017 annual convention of the American Educational Research Association in Chicago, IL.

Correspondence concerning this article should be addressed to James E. Pustejovsky, University of Texas at Austin, Department of Educational Psychology, 1912 Speedway, Stop D5800, Austin, TX 78712-1289. Email: pusto@austin.utexas.edu.

Abstract

In linear regression models estimated by ordinary least squares, it is often desirable to use hypothesis tests and confidence intervals that remain valid in the presence of heteroskedastic errors. Wald tests based on heteroskedasticity-consistent covariance matrix estimators are a well known and widely used method that remains asymptotically valid under heteroskedasticity of an unspecified form. Wald-type t-tests based on HCCMEs maintain nominal rejection rates when the sample size is large, but they are not always accurate with small samples, and it can be difficult to determine whether a given sample is large enough to trust the asymptotic approximation. This paper reviews several approaches to approximating the sampling distribution of HCCME t-tests and thereby improving the accuracy of rejection rates in small samples. Using simulations, we investigate the relative performance of Satterthwaite, Edgeworth, and saddlepoint approximations under several data generating processes. We find that certain distributional approximations perform comparably to or better than conventional tests based on HCCMEs, with marked improvements when the nominal Type-I error rate is less than .05.

Keywords: heteroskedasticity; sandwich estimator; robust covariance estimator; linear regression; Satterthwaite approximation; saddlepoint approximation; Edgeworth approximation

Heteroskedasticity-robust tests in linear regression: A review and evaluation of small-sample corrections

Introduction

Linear regression models, estimated by ordinary least squares (OLS), are one of the most important and ubiquitous tools in applied statistical work. Classical hypothesis tests and confidence intervals for linear regression coefficients rely on the assumption that the model errors are homoskedastic, or have constant variance for all values of the covariates. In practice though, it can be difficult to diagnose violations of this assumption, and similarly difficult to construct and defend other assumptions about how the error variances relate to the covariates. Thus, it is often desirable to use methods that remain valid under models with heteroskedasticity of an unknown form.

A well-known approach to inference in this setting is based on heteroskedasticity-consistent covariance matrix estimators (HCCMEs), which yield asymptotically consistent estimates of the sampling variance of OLS coefficient estimates under quite general conditions (Eicker, 1967; Huber, 1967; White, 1980). HCCMEs are an attractive tool because they rely on weaker assumptions than classical methods. However, they also have the drawback that it is not always clear whether a given sample is sufficently large to trust the asymptotic approximations by which they are warranted. Furthermore, when the sample size is small, it is known that some of the HCCMEs tend to be too liberal, producing variance estimates that are biased towards zero and hypothesis tests with greater than nominal size (Long & Ervin, 2000).

Since White (1980) introduced the HCCME in econometrics, methods for improving the finite-sample properties of HCCMEs have been studied extensively. The most well-known strand of this work has considered modified forms of the HCCME that produce more accurate tests and CIs in finite samples. MacKinnon and White (1985) and Davidson and MacKinnon (1993) proposed several such modifications that are now readily available in software. Based upon an extensive set of simulations, Long and Ervin (2000) demonstrated that one of these modifications, known as HC3, performs substantially better than the others. As a result, HC3 is the default in software such as the R package `sandwich` (Zeileis, 2004), although White's original HCCME remains the default in SAS `proc reg` and Stata's `regress` command with `vce(robust)`. More recently, several further variations on the HCCMEs have been proposed (Cribari-Neto, 2004; Cribari-Neto & da Silva, 2011; Cribari-Neto, Souza, & Vasconcellos, 2007), which aim to improve upon the performance of HC3 in models where the regressors exhibit high leverage. For hypothesis testing, HCCMEs are typically used to calculate t-statistics, which are compared to standard normal or $t(n - p)$ reference distributions, where n is the sample size and p is the dimension of the coefficient vector.

An alternative approach to improving the small-sample properties of hypothesis tests based on HCCMEs is to find a better approximation to the null sampling distribution of the test statistic. Several such approximations have been proposed, including Satterthwaite (Lipsitz, Ibrahim, & Parzen, 1999), Edgeworth (Kauermann & Carroll, 2001; Rothenberg, 1988), and saddlepoint approximations (McCaffrey & Bell, 2006). Although there is evidence that each of these approximations improves upon the standard, large-sample tests, their finite-sample performance has been studied only under a very limited range of data-generating processes. Also, existing simulation evidence about these tests has focused almost exclusively on Type-I error rates at the nominal $\alpha = .05$ level. Furthermore, it appears that the existing approximations have been developed in isolation, without reference to previous work, and they have received little subsequent attention (e.g., none are discussed in a recent review of heteroskedasticity-robust inference by MacKinnon, 2013). In contrast to the various HC corrections, none of the distributional approximations are implemented in standard software packages for data analysis.

In this paper, we review the various small-sample approximations for hypothesis tests based on HCCMEs, using a common notation in order to facilitate comparisons among them. In so doing, we identify several further variations on the approximations that have not previously been considered. We then evaluate the performance of these approximations, along with the standard methods, in a simulation study. Though not comprehensive, our simulations include several different distributional models, varying degrees of heteroskedasticity, varying degrees of leverage (which is known to have a strong influence on the performance of HCCMEs), and several nominal Type-I error rates, including levels lower than $\alpha = .05$. Lower α levels are of interest because scholars have recently argued for reducing the conventional $\alpha = .05$ level to level of .005 (Benjamin et al., 2018). Even under current conventions, lower α levels are important for analyses involving multiple comparisons corrections

HCCMEs are a special case of the general class of cluster-robust covariance matrix estimators (CRCMEs), also known as sandwich estimators or linearization estimators, which are commonly used in regression analysis of multi-stage survey data (Fuller, 1975; Skinner, 1989), econometric panel data models (Arellano, 1987; White, 1984), and generalized estimating equations for longitudinal data (Liang & Zeger, 1986). CRCMEs are useful for variance estimation in settings where the error structure is both heteroskedastic and dependent within clusters of observations. Some of the small-sample tests considered in this paper were developed for CRCMEs (i.e., Bell & McCaffrey, 2002; McCaffrey & Bell, 2006), while the others are readily extended to this more general case. We focus on the case of heteroskedastic (but not clustered) linear regression for sake of clarity and in order to keep the simulation studies tractable. Furthermore, the similarity of HCCMEs and CRCMEs suggests that our findings will provide direction regarding which small-sample methods will perform well in the more general case.

The remainder of the paper proceeds as follows. In Section , we lay out the model, define the HCCMEs, and outline relevant distribution theory. Section reviews several small-sample corrections for hypothesis tests based on HCCMEs. Section reports simulation evidence on the size properties of these tests. Section discusses limitations and future research directions.

Theoretical context

Model and notation

We shall consider the regression model

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \epsilon_i, \tag{1}$$

for $i = 1, ..., n$, where y_i is the outcome, $\mathbf{x}_i$ is a $1 \times p$ row-vector of covariates (including an intercept) for observation i, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and ϵ_i is a mean-zero error term with variance σ_i^2. We shall assume that the errors are mutually independent. Let $\mathbf{y} = (y_1, ..., y_n)'$ denote the $n \times 1$ vector of outcomes, $\mathbf{X} = (\mathbf{x}_1', ..., \mathbf{x}_n')'$ be the $n \times p$ design matrix, and $\boldsymbol{\epsilon}$ be the $n \times 1$ vector of errors with $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, ..., \sigma_n^2)$. Let $\mathbf{M} = (\mathbf{X}'\mathbf{X}/n)^{-1}$. The OLS estimator of $\boldsymbol{\beta}$ is then given by

$$\hat{\boldsymbol{\beta}} = \mathbf{M}\mathbf{X}'\mathbf{y}/n. \tag{2}$$

Let $\mathbf{I}$ denote an $n \times n$ identity matrix and $\mathbf{H} = \mathbf{X}\mathbf{M}\mathbf{X}'/n$ denote the hat matrix, which has entries $h_{ij} = \mathbf{x}_i \mathbf{M} \mathbf{x}_j'/n$. Finally, let $e_i = y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}$, $i = 1, ..., n$ denote the residuals.

In what follows, the aim will be to test a hypothesis regarding a linear combination of the regression coefficients, expressed as $H_0 : \mathbf{c}'\boldsymbol{\beta} = k$, with target Type-I error rate α. All tests under consideration are based on the Wald statistic

$$T = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}} - k}{\sqrt{V}}, \tag{3}$$

where V is some estimator for $\mathrm{Var}\left(\mathbf{c}'\hat{\boldsymbol{\beta}}\right)$. In what follows, we shall use superscripts on T that correspond to the superscript for the variance estimator used to calculate it.

If the errors are homoskedastic, so that $\sigma_i^2 = \sigma^2$ for $i = 1, ..., n$, then the hypothesis can be evaluated using the standard t test. The variance of $\mathbf{c}'\boldsymbol{\beta}$ is then estimated by $V^{hom} = \hat{\sigma}^2 \mathbf{c}'\mathbf{M}\mathbf{c}/n$, where $\hat{\sigma}^2 = \sum_{i=1}^n e_i^2/(n-p)$. Under H_0 and assuming that the errors are normally distributed, the test statistic follows a t distribution with $n - p$ degrees of freedom. Thus, H_0 is rejected if $\left|T^{hom}\right| > F_t^{-1}\left(1 - \frac{\alpha}{2}; n - p\right)$, where $F_t^{-1}(x; \nu)$ is the quantile function for a t distribution with ν degrees of freedom. If the errors are instead heteroskedastic, the variance estimator V^{hom} will be inconsistent and this t test will not generally have correct size.

HCCMEs

Allowing for heteroskedasticity, the true variance of the OLS estimator is

$$\text{Var}\left(\mathbf{c}'\hat{\boldsymbol{\beta}}\right) = \frac{1}{n}\mathbf{c}'\mathbf{M}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\mathbf{x}_i\mathbf{x}_i'\right)\mathbf{M}\mathbf{c} \tag{4}$$

The HCCCMEs estimate $\text{Var}\left(\mathbf{c}'\hat{\boldsymbol{\beta}}\right)$ by replacing the σ_i^2 with estimates based on the squared residuals. All of the HCCMEs have the same general form

$$V^{HC} = \frac{1}{n}\mathbf{c}'\mathbf{M}\left(\frac{1}{n}\sum_{i=1}^{n}\omega_i e_i^2\mathbf{x}_i\mathbf{x}'\right)\mathbf{M}\mathbf{c} \tag{5}$$

where $\omega_1, ..., \omega_n$ are weighting terms that differ for the various HC estimators. Under general assumptions, the weak law of large numbers ensures that the middle term in Equation (5) converges to the corresponding term in (4) as the sample size increases, so that V^{HC} is asymptotically consistent (White, 1980).

White (1980) originally described the HCCME without any correction factor, which is equivalent to taking $\omega_i = 1$ for $i = 1, ..., n$. This form has come to be known as HC0. Subsequently, various correction factors have been proposed that aim to improve on the finite-sample behavior of HC0. Following common convention, we refer to these correction factors by number. Their forms are as follows:

HC1:	$\omega_i = n/(n-p)$	
HC2:	$\omega_i = (1-h_{ii})^{-1}$	
HC3:	$\omega_i = (1-h_{ii})^{-2}$	
HC4:	$\omega_i = (1-h_{ii})^{-\delta_i}$,	$\delta_i = \min\{h_{ii}n/p, 4\}$
HC4m:	$\omega_i = (1-h_{ii})^{-\delta_i}$,	$\delta_i = \min\left\{h_{ii}n/p, 1\right\} + \min\left\{h_{ii}n/p, 1.5\right\}$
HC5:	$\omega_i = (1-h_{ii})^{-\delta_i}$,	$\delta_i = \min\left\{h_{ii}n/p, \max\left\{4, 0.7h_{(n)(n)}n/p\right\}\right\}$

MacKinnon and White (1985) suggested HC1, which uses an ad hoc correction similar to the correction used for $\hat{\sigma}^2$, and HC2, which has the property that V^{HC2} is exactly unbiased when the errors are homoskedastic. Davidson and MacKinnon (1993) proposed HC3 as an approximation to the leave-on-out jackknife variance estimator.

Cribari-Neto and colleagues subsequently proposed three further variations, HC4 (Cribari-Neto, 2004), HC4m (Cribari-Neto & da Silva, 2011), and HC5 (Cribari-Neto et al., 2007), all of which aim to improve upon HC3 for design matrices where some observations have high leverage. All of these correction factors inflate the squared residual term to a greater extent when an observation has a higher degree of leverage. HC4 truncates the degree of inflation at 4 times the average leverage. Compared to HC4, HC4m inflates observations with lower leverage more strongly, while truncating the maximum degree of inflation at 2.5 times the average. In HC5, the truncation point depends on the maximum leverage value and will exceed 4 if the maximum leverage value is more than 5.71 times the average.

For any of the HCCMEs and under weak conditions on the error and covariate distributions, the robust Wald statistic $T^{HC} = \left(\mathbf{c}'\hat{\boldsymbol{\beta}} - k\right)/\sqrt{V^{HC}}$ converges in distribution to $N(0,1)$ as n increases to infinity. Thus, an asymptotically correct test can be constructed by rejecting H_0 when $\left|T^{HC}\right|$ is greater than the $1-\alpha/2$ critical value from a standard normal distribution. In practice, it is common to instead use the critical value from a t distribution with $n-p$ degrees of freedom. However, use of the t_{n-p} reference distribution is only an ad hoc approximation. In Section , we review several distinct, more theoretically grounded approximations to the null sampling distribution of T^{HC}.

Distribution of V^{HC}

The approximations described in the following section all involve expressions for the distribution of V^{HC}. Thus, we first briefly summarize the relevant distribution theory.

For any of the correction factors (HC0-HC5), the variance estimator V^{HC} is a quadratic form in the residuals (and thus also in the errors), which can be written as

$$V^{HC} = \sum_{i=1}^{n} \omega_i \left(g_i e_i\right)^2 = \mathbf{e}'\mathbf{A}\mathbf{e} = \boldsymbol{\epsilon}'\mathbf{B}\boldsymbol{\epsilon},$$

where $g_i = \mathbf{x}_i \mathbf{M}\mathbf{c}/n$, $\mathbf{A} = \text{diag}\left(\omega_1 g_1^2, ..., \omega_n g_n^2\right)$, and $\mathbf{B} = \left(\mathbf{I} - \mathbf{H}\right)\mathbf{A}\left(\mathbf{I} - \mathbf{H}\right)$ (Bell & McCaffrey, 2002; Cribari-Neto & da Silva, 2011). It follows from the properties of quadratic forms that

$$\text{E}\left(V^{HC}\right) = \text{tr}\left[\mathbf{B}\boldsymbol{\Sigma}\right]. \tag{6}$$

Furthermore, assuming that the model errors are normally distributed, the variance of the quadratic form is

$$\text{Var}\left(V^{HC}\right) = 2\text{tr}\left[\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}\boldsymbol{\Sigma}\right] = 2\text{tr}\left[\mathbf{B}\left(\mathbf{B} \circ \mathbf{S}\right)\right], \tag{7}$$

where $\circ$ denotes the element-wise (Hadamard) product and $\mathbf{S}$ has entries $S_{ij} = \sigma_i^2 \sigma_j^2$ (Lipsitz et al., 1999).

Again assuming that the model errors are normally distributed, the sampling distribution of V^{HC} can be expressed as a weighted sum of χ_1^2 random variables. Note that the matrix $\mathbf{B}\boldsymbol{\Sigma}$ has rank $n - p$. Let $\lambda_1, ..., \lambda_{n-p}$ denote its non-zero eigenvalues, arranged in descending order. Let $Z_1, ..., Z_{n-p}$ denote independent χ_1^2 random variates. Then the distribution

$$V^{HC} \stackrel{d}{=} \sum_{i=1}^{n-p} \lambda_i Z_i, \tag{8}$$

where $\stackrel{d}{=}$ means that two quantities have identical distributions (Mathai & Provost, 1992, Eq. 4.1.1).

Distributional approximations

This section reviews four approximations to the null sampling distribution of T^{HC}, including a Satterthwaite approximation, two different Edgeworth-type approximations, and a saddlepoint approximation. As will be seen, all of the approximations involve quantities that depend on the unknown error variances. Thus, a key consideration in developing feasible versions of these approximations is how to estimate the error variances.

Past proposals have each considered different strategies, including estimating the errors empirically (as in the HCCME itself) or by assuming that they follow a known structure.

Satterthwaite approximation

Lipsitz et al. (1999) proposed a hypothesis testing procedure that is based on a Satterthwaite approximation for the distribution of T^{HC}, where V^{HC} is calculated using the HC2 form of the variance estimator. In this approach, the distribution of V^{HC} is approximated by a multiple of a χ^2_ν distribution, with degrees of freedom chosen to match the first two moments of V^{HC} (Satterthwaite, 1946). In the abstract, the Satterthwaite degrees of freedom are given by

$$\nu = 2\left[\mathrm{E}\left(V^{HC}\right)\right]^2 / \mathrm{Var}\left(V^{HC}\right).$$

With these degrees of freedom, the null hypothesis is rejected if $|T^{HC}| > F_t^{-1}(1-\alpha/2, \nu)$. Readers may be familiar with Satterthwaite approximation because it is the basis of the degrees of freedom commonly used in the two-sample t-test assuming unequal variances (Welch, 1947).

The mean and variance of V^{HC} involve the unknown error variances $\mathbf{\Sigma}$, and so must be estimated in order to calculate the Satterthwaite approximation. Lipsitz and colleagues proposed to use V^{HC} as an estimate of its own expectation and to estimate $\mathrm{Var}\left(V^{HC}\right)$ based on the model residuals. Specifically, let $\hat{\mathbf{S}}$ be the matrix with entries

$$\hat{S}_{ii} = \frac{1}{3}\omega_i^2 e_i^4 \quad \text{for } i = 1, ..., n \quad \text{and} \quad \hat{S}_{ij} = \frac{\omega_i \omega_j e_i^2 e_j^2}{2\omega_i\omega_j h_{ij}^2 + 1} \quad \text{for } i \neq j,$$

to be used as an estimate of $\mathbf{S}$ in Equation (7). The empirically estimated degrees of freedom are then given by

$$\nu_E = \frac{\left(V^{HC}\right)^2}{\mathrm{tr}\left[\mathbf{B}\left(\mathbf{B} \circ \hat{\mathbf{S}}\right)\right]}. \tag{9}$$

Bell and McCaffrey (2002) proposed a similar test (also based on a Satterthwaite approximation) for regression coefficients with standard errors estimated by a CRCME. Rather than estimate the moments of V^{HC} empirically, Bell and McCaffrey (2002) suggested calculating (6) and (7) based on a working model for the error structure (see also Imbens & Kolesar, 2015). In the present context, a leading candidate for a working model is to assume that the errors are homoskedastic, so that $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$. The degrees of freedom then reduce to

$$\nu_M = \frac{\left(\sum_{i=1}^{n}(1 - h_{ii})\omega_i g_i^2\right)^2}{\sum_{i=1}^{n}(1 - h_{ii})^2 \omega_i^2 g_i^4 + \sum_{i=1}^{n}\sum_{j \neq i} h_{ij}^2 \omega_i \omega_j g_i^2 g_j^2}. \tag{10}$$

In principle, these "model-based" degrees of freedom could be used with any of the HC estimators; in practice, however, the HC2 estimator is a natural choice because it is exactly unbiased under homoskedasticity. Using the HC2 correction factors, the degrees of freedom simplify further to

$$\nu_M = \frac{\left(\sum_{i=1}^{n} g_i^2\right)^2}{\sum_{i=1}^{n} g_i^4 + \sum_{i=1}^{n}\sum_{j \neq i} \frac{g_i^2 g_j^2 h_{ij}^2}{(1 - h_{ii})(1 - h_{jj})}} \tag{11}$$

(cf. Kauermann & Carroll, 2001, Eq. 5).

Kauermann and Carroll's Edgeworth approximation

Kauermann and Carroll (2001) proposed approximate confidence intervals for $\mathbf{c}'\hat{\boldsymbol{\beta}}$ based on an Edgeworth approximation to the distribution of T^{HC}. Their approximation is based on the assumption that V^{HC} is unbiased and independent of $\mathbf{c}'\hat{\boldsymbol{\beta}}$. Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal cumulative distribution function and density function and let $z_\alpha = \Phi^{-1}(1 - \alpha/2)$ denote the $1 - \alpha/2$ critical value. The hypothesis testing procedure corresponding to the confidence interval proposed by Kauermann and Carroll (2001) rejects the null if $\left|T^{HC}\right| > z_{\tilde{\alpha}}$, where $\tilde{\alpha}$ is defined implicitly as the solution to

$$\alpha = \tilde{\alpha} + \frac{\phi\left(z_{\tilde{\alpha}}\right)}{2\nu}\left(z_{\tilde{\alpha}}^3 + z_{\tilde{\alpha}}\right). \tag{12}$$

Equivalently, the p-value for the test is given by

$$p = 2\left[1 - \Phi\left(\left|T^{HC}\right|\right)\right] + \frac{\phi\left(\left|T^{HC}\right|\right)}{2\nu}\left(\left|T^{HC}\right|^3 + \left|T^{HC}\right|\right). \tag{13}$$

Kauermann and Carroll focus on the HC2 variance estimator and calculate its degrees of freedom based on the working assumption that the errors are actually homoskedastic, as in ν_M from Equation (11). An alternative would be to use the empirical degrees of freedom estimate, ν_E, from Equation(9).

Kauermann and Carroll (2001) also offer the following further approximation for the critical value $z_{\tilde{\alpha}}$:

$$z_{\tilde{\alpha}} = F_t^{-1}\left(1 - \frac{\alpha}{2}; n-p\right) + \frac{z_\alpha^3 + z_\alpha}{4\nu} - \frac{\left(z_\alpha^3 + z_\alpha\right)\left(\sum_{i=1}^n g_i^2\right)^2}{4(n-p)}. \tag{14}$$

This further approximation is convenient for calculating a confidence interval for $\mathbf{c}'\hat{\boldsymbol{\beta}}$ because it avoids the need to numerically solve Equation (12). The simulation studies reported in the following section evaluate both the p-value approximation and the CI approximation.

Rothenberg's Edgeworth approximation

Prior to Kauermann and Carroll (2001), Rothenberg (1988) developed an Edgeworth approximation for the distribution of T^{HC}, calculated using the HC0 variance estimator. Rothenberg's approximation differs from Kauermann and Carroll's in two key ways. First, it allows for the possibility that V^{HC} is a biased estimator of $\text{Var}\left(\mathbf{c}'\hat{\boldsymbol{\beta}}\right)$; such will be the case for V^{HC0} if the errors are homoskedastic, for instance. Second, it allows for the possibility of dependence between $\mathbf{c}'\hat{\boldsymbol{\beta}}$ and V^{HC}, which arises when the errors are *not* homoskedastic.

Although originally developed for the HC0 estimator, Rothenberg's approximation is readily applied to the other forms too; we give the general version here. Let

$$\begin{aligned}
f_i &= g_i\sigma_i^2 - \sum_{j=1}^n h_{ij} g_j \sigma_j^2 \\
q_i &= \left(\sum_{j=1}^n h_{ij}^2 \sigma_j^2\right) - 2h_{ii}\sigma_i^2 \\
a &= \left(\sum_{i=1}^n \omega_i g_i^2 f_i^2\right)\left(\sum_{i=1}^n g_i^2\sigma_i^2\right)^{-2} \\
b &= \left(\sum_{i=1}^n \omega_i g_i^2 q_i\right)\left(\sum_{i=1}^n g_i^2\sigma_i^2\right)^{-1}
\end{aligned}$$

Rothenberg's Edgeworth approximation is then given by

$$\Pr\left(T^{HC} \leq t\right) \approx \Phi\left[t\left(1 - \frac{1+t^2}{4\nu} + \frac{a\left(t^2-1\right)+b}{2}\right)\right].$$

Here, the a term measures covariance between $\mathbf{c}'\hat{\boldsymbol{\beta}}$ and V^{HC}; the b term measures the relative bias of V^{HC}; and ν is the Satterthwaite degrees of freedom.

Based on this Edgeworth approximation, Rothenberg (1988) proposed a test in which the null hypothesis is rejected if $\left|T^{HC}\right| > t_\alpha$, where the critical value t_α is defined by

$$t_\alpha = z_\alpha\left(1 + \frac{z_\alpha^2+1}{4\nu} - \frac{a\left(z_\alpha^2-1\right)+b}{2}\right). \tag{15}$$

It can be seen that this critical value is similar to Kauermann and Carroll's closed-form approximate critical value from Equation (14), the only differences being that the first term uses a standard normal quantile rather than a t_{n-p} quantile and that the third terms differ.

In practice, the a and b terms and the degrees of freedom must be estimated because they depend on the unknown error variances. Rothenberg proposed to do so by replacing values of σ_i^2 with e_i^2 in the expressions for a and b and using $\nu_R = \left(\sum_{i=1}^n g_i^2 e_i^2\right)^2 \left(\sum_{i=1}^n g_i^4 e_i^4/3\right)^{-1}$ as an empirical degrees of freedom approximation. For purposes of simplicity, the simulation studies described in the next section use ν_E instead of ν_R. An alternative approach—not considered by Rothenberg—is to calculate a, b, and ν based on the assumption that the errors are homoskedastic. In this case, $a = 0$, $b = -\left(\sum_{i=1}^n h_{ii}\omega_i g_i^2\right) / \left(\sum_{i=1}^n g_i^2\right)$, and the degrees of freedom are equal to ν_M. Using the “model-based” estimates of the adjustment quantities may be quite reasonable, considering that if the bias of V^{HC} could be well-estimated empirically, one could simply correct the estimator itself.

Saddlepoint approximation

McCaffrey and Bell (2006) developed small-sample adjustments to test statistics based on CRCMEs, of which the HC estimators are a special case. They considered both a Satterthwaite approximation (similar to Lipsitz et al.) and a saddlepoint approximation for the distribution of the test statistic, finding that the latter produced tests with more accurate size.

The saddlepoint technique is a tool for approximating the density or distribution of a random variable based on its cumulant generating function (Goutis & Casella, 1999; Huzurbazar, 1999). The test proposed by McCaffrey and Bell (2006) is derived by first representing $\left|T^{HC}\right|$ as a ratio of weighted sums of independent χ^2_1 variates, then approximating its cumulative distribution using a saddlepoint formula due to Lugannani and Rice (1980). The cumulative distribution of T^{HC} can be expressed as

$$\Pr\left(\left|T^{HC}\right| \leq t\right) = \Pr\left(\frac{\left(\mathbf{c}\hat{\boldsymbol\beta} - k\right)^2}{\text{Var}\left(\mathbf{c}\hat{\boldsymbol\beta}\right)} - t^2 \frac{V^{HC}}{\text{Var}\left(\mathbf{c}\hat{\boldsymbol\beta}\right)} \leq 0\right).$$

Observe that $\left(\mathbf{c}\hat{\boldsymbol\beta} - k\right)^2 / \text{Var}\left(\mathbf{c}\hat{\boldsymbol\beta}\right) \sim \chi^2_1$ and that V^{HC} is distributed as a weighted sum of χ^2_1 random variables, as in Equation (8). McCaffrey and Bell (2006) assume that V^{HC} is unbiased, so that

$$\text{E}\left(V^{HC}\right) = \text{tr}\left(\mathbf{B\Sigma}\right) = \sum_{j=1}^{n-p} \lambda_j,$$

and that $\hat{\boldsymbol\beta}$ is independent of V^{HC}. It then follows that the $\Pr\left(\left|T^{HC}\right| \leq t\right)$ can be expressed as $\Pr(Z \leq 0)$, where $Z = \sum_{i=0}^{n-p} \gamma_i Z_i$, $\gamma_0 = 1$, $\gamma_i = -t^2 \lambda_i / \sum_{j=1}^{n-p} \lambda_j$ for $i = 1, ..., n-p$, and $Z_0, ..., Z_{n-p} \stackrel{\text{iid}}{\sim} \chi^2_1$.

The saddlepoint approximation for $\Pr(Z \leq 0)$ is obtained as follows. Let s be the saddlepoint, defined implicitly as the solution to

$$\sum_{i=0}^{n-p} \frac{\gamma_i}{1 - 2\gamma_i s} = 0.$$

The saddlepoint must be calculated numerically (e.g., via a grid search).[1] Define the quantities r and q as

$$r = \text{sign}(s)\sqrt{\sum_{i=0}^{n-p} \log\left(1 - 2\gamma_i s\right)}, \qquad q = s\sqrt{2\sum_{i=0}^{n-p} \frac{\gamma_i^2}{\left(1 - 2\gamma_i s\right)^2}}.$$

Then

$$\Pr(Z \leq 0) \approx \begin{cases} \Phi(r) + \phi(r)\left[\frac{1}{r} - \frac{1}{q}\right] & s \neq 0 \\ \frac{1}{2} + \frac{\sum_{i=0}^{n-p} \gamma_i^3}{3\sqrt{\pi}\left(\sum_{i=0}^{n-p} \gamma_i^2\right)^{3/2}} & s = 0 \end{cases} \tag{16}$$

(Lugannani & Rice, 1980). Given an observed value for the t-statistic t^{HC}, a p-value for H_0 can be calculated by taking $\gamma_i = -\left(t^{HC}\right)^2 \lambda_i / \sum_{j=1}^n \lambda_j$ for $i = 1, ..., n - p$, finding s, r, and q, and evaluating $1 - \Pr(Z \leq 0)$ using Equation (16). In order to avoid numerical inaccuracy, we evaluate the saddlepoint using the second line of Equation (16) if $|s| < .01$.

In practice, the unknown error variances must be estimated in order to find the eigenvalues of $\mathbf{B\Sigma}$. McCaffrey and Bell (2006) propose to do so based on a working model. For instance, assuming that the errors are homoskedastic implies that the eigenvalues of $\mathbf{B}$ may be used in the saddlepoint calculations. An alternative, not considered by McCaffrey and Bell (2006), would be to use the eigenvalues of $\mathbf{B\hat{\Sigma}}$, where $\hat{\mathbf{\Sigma}} = \text{diag}\left(e_1^2, ..., e_n^2\right)$. The simulation studies examine the performance of both the working model approach and the empirical approach to calculating the saddlepoint approximation, analogous to using either the empirical or model-based degrees of freedom in conjunction with the other approximations.

Remarks

We have reviewed several approximations for the null sampling distribution of T^{HC} and have also noted that any of the approximations could be applied using either empirical

[1] For programming, it is helpful to note that $(2\gamma_1)^{-1} < s < 0$ if $|T^{HC}| < 1$; $0 < s < 1/2$ if $|T^{HC}| > 1$; and $s = 0$ if $|T^{HC}| = 1$.

estimates of the model errors or estimates based on an assumed working model, such as homoskedasticity. All of the approximations are derived under the assumption that the model errors are normally distributed, and several of them invoke the additional assumption that V^{HC} is independent of the OLS coefficient estimator, which will not hold precisely unless the errors are homoskedastic. The approximations may differ in the extent to which their performance suffers under data-generating models with non-normal or heteroskedastic errors. Furthermore, some versions of the approximations involve a working model, and it is not immediately apparent how discrepancies between the working model and the true data generating model will affect their performance. Thus, it is not obvious on the basis of their derivations alone which approach is most accurate with small samples, nor whether any of the approaches represents an improvement on more conventional practice (i.e., using T^{HC3} with a $t(n-p)$ reference distribution).

Simulations

This section reports a large simulation study that investigate the performance the distributional approximations under a range of conditions, including conditions in which errors are non-normally distributed.

Simulation design

To keep the dimension of the simulation manageable, we examine test performance under a model with a single regressor. It is known that the performance of conventional tests based on HCCMEs are influenced not only by sample size, but by the distribution of the regressors (Chesher & Austin, 1991; Cribari-Neto, 2004; Kauermann & Carroll, 2001). Specifically, observations with high leverage tend to distort the size of the conventional tests. In order to study the performance of HCCME-based tests under varying degrees of leverage, we simulated the regressor from a χ^2 distribution with degrees of freedom selected to control the skewness:

$$X_1, ..., X_n \stackrel{\text{iid}}{\sim} \frac{\gamma^2 \chi^2_{8/\gamma^2} - 8}{4\gamma}.$$

The distribution of X therefore has mean zero, unit variance, and skewness γ. We then simulated the outcome as

$$Y_i = \beta X_i + \sigma_i \epsilon_i, \quad i = 1, ..., n.$$

where the errors $\epsilon_1, ..., \epsilon_n$ were simulated from one of three different error distributions, including the standard normal, t_5 (scaled to have unit variance), or χ^2_5 (centered and scaled) distribution, and the skedasticity function was taken to be $\sigma_i = \exp(\zeta X_i)$. The constant ζ controls the degree of heteroskedasticity, with $\zeta = 0$ corresponding to homoskedasticity and $\zeta = 0.2$ corresponding to substantial heteroskedasticity.

Based on this model, we simulated samples of 25, 50, or 100, using $\gamma = \frac{1}{2}, 1, 2$, and $\zeta = 0, ..., 0.2$ in steps of 0.02. For each simulated dataset, we tested the hypothesis $\beta = 0$ using 17 different procedures. First, we calculated tests based on the HC0, HC1, HC2, HC3, HC4, HC4m, and HC5 adjustment factors compared to conventional $t(n - p)$ critical values. Second, we calculated tests using the Satterthwaite approximation (with HC2), both Edgeworth approximations from Kauermann and Carroll (2001, also using HC2), the Rothenberg (1988) Edgeworth approximation (with HC0), and the saddlepoint approximation (with HC2). For each of the distributional approximations, we examined both empirical- and model-based versions of the correction. We considered nominal type-I error levels of $\alpha = .005$, .010, and .050. Empirical rejection rates are estimated from 5×10^4 replications.

Results: Size

Due to space constraints, we present only selected results, focusing initially on the sample size of $n = 50$. We omit results for the t_5 error distribution, which ares very similar to the results for normal errors. The supplementary materials provide complete numerical results for all conditions, as well as R code for replicating all calculations.

We first consider the empirical size of all seventeen tests. Figure 1 depicts the rejection rates of the conventional t-tests based on the HC3, HC4, HC4m, and HC5 estimators, at a sample size of $n = 50$, as a function of the degree of heteroskedasticity.

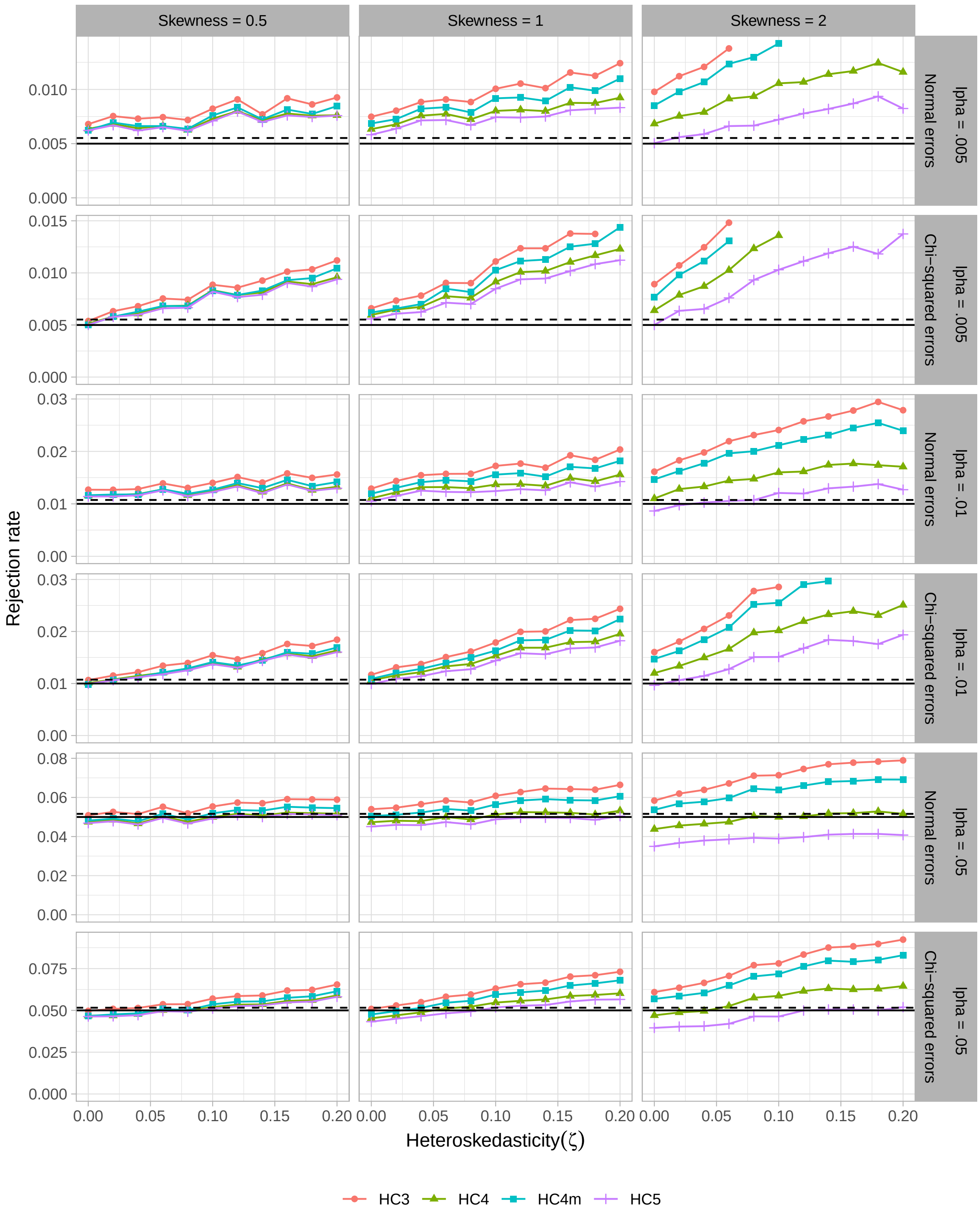

Figure 1. Rejection rates of conventional tests based on HC3, HC4, HC4m, and HC5 for $n = 50$. The solid horizontal line indicates the stated α level and the dashed line indicates an upper confidence bound on simulation error.

Each plot is repeated for varying degrees of skewness (columns) and nominal level (rows), and with either normal of χ^2_5 errors (rows). In order to preserve detail, rejection rates higher than 3α are omitted from this and the following figures. Tests based on HC0, HC1, and HC2 have strictly higher rejection rates than HC3, and so we omit them from Figure 1.

It can be seen that the test with HC3 does not maintain the nominal level except when the degree of heteroskedascity is small and skewness is mild. Its performance also degrades at smaller values of α. The test with HC5 behaves similarly to that with HC3, but has even higher rejection rates when skewness is less severe.

Of the conventional tests, using HC4 produces rejection rates that are closest to nominal across the conditions examined. At $\alpha = .05$, its rejection rates remain very close to nominal when errors are normal, although they are slightly larger than nominal when errors are χ^2_5 and more strongly heteroskedastic. At lower values of α, however, even HC4 has higher-than-nominal rejection rates, and these are more strongly affected by the degree of heteroskedasticity. Finally, rejection rates of the conventional test with HC4m are intermediate between those of HC4 and those of HC5. In further analysis, we focus only on the test with HC4 because its rejection rates are consistently closer to the nominal level across all conditions.

Figure 2 depicts the rejection rates of the tests involving Edgeworth approximations, again at a sample size of $n = 50$; its construction follows that of Figure 1. The tests include Kauermann and Carroll's (2001) confidence interval approximations (denoted as KCCI_E for the empirical version and KCCI_H for the version derived under homoskedasticity) and p-value approximations (denoted as KCp_E for the empirical version and KCp_H for the version derived under homoskedasticity), as well as Rothenberg's (1988) approximation derived under homoskedasticity (denoted as RCI_H). The empirical version of Rothenberg's approximation is omitted because its rejection rates were far in excess of nominal, even at the largest sample size of $n = 100$.

All of the Edgeworth approximations perform well at the $\alpha = .05$ level under the

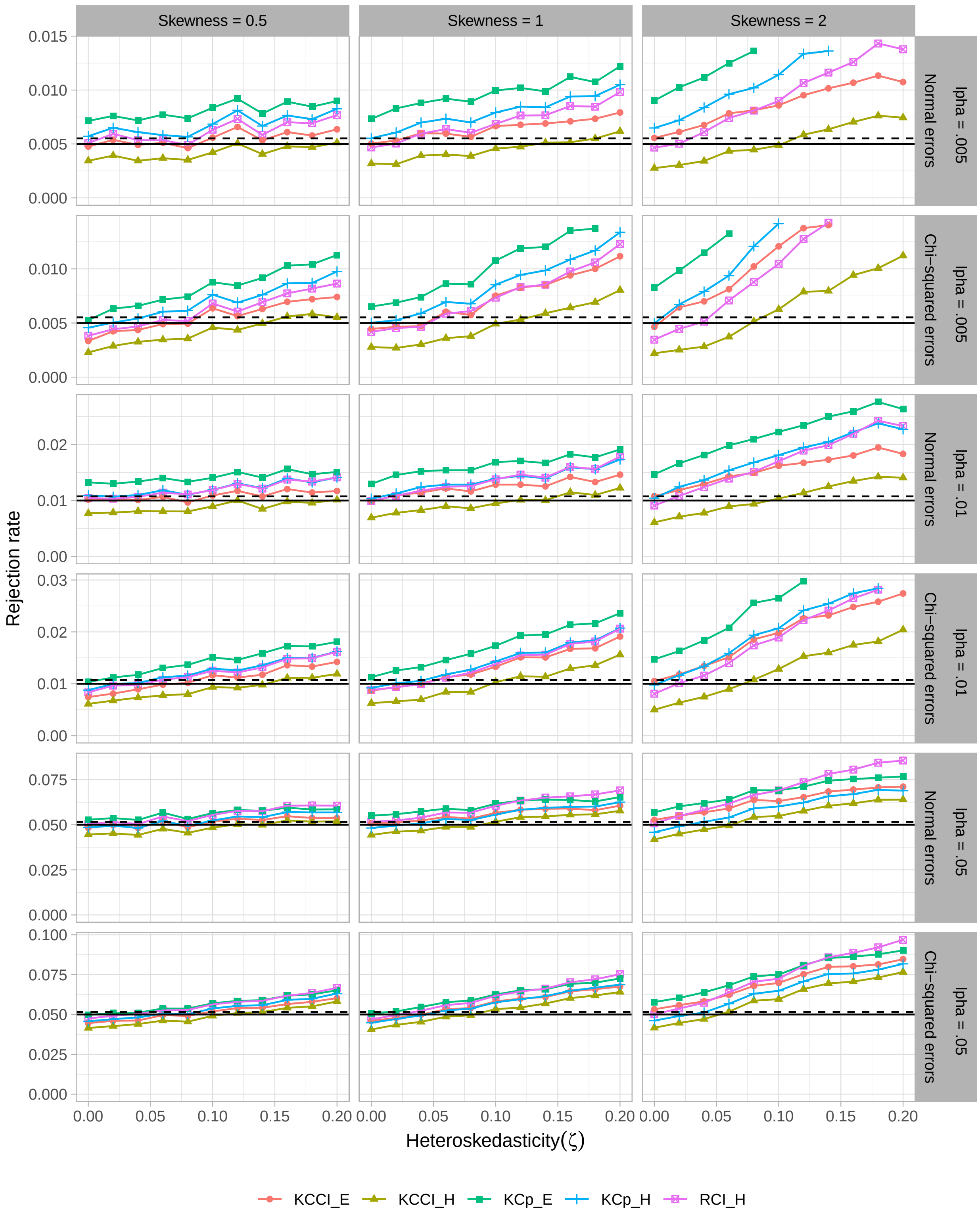

Figure 2. Rejection rates of Edgeworth approximation tests for $n = 50$. The solid horizontal line indicates the stated α level and the dashed line indicates an upper confidence bound on simulation error.

model with normal errors and small skewness. As with the conventional tests, the rejection rates of all of the Edgeworth approximations increase when the covariate is strongly skewed and as the degree of heteroskedasticity grows. Rejection rates of the empirical and homoskedastic p-value approximations generally tend to exceed the rates of the confidence interval approximations. They also exceed the nominal α, particularly under models with a highly skewed covariate and the smaller values of α. The rejection rates of Kauermann and Carroll's homoskedastic confidence interval approximation are closest to maintaining the nominal level. We therefore focus on it in subsequent analysis.

Following the same layout as in previous figures, Figure 3 depicts the rejection rates of the empirical and homoskedastic versions of the saddlepoint and Satterthwaite approximation tests. It can be seen that the empirical saddlepoint test tends to have lower-than-nominal rejection rates when the covariate skewness is small or moderate, and this holds across nominal α levels and error distributions. However, under high skewness its rejection rates exceed nominal under some condition—particularly for strong heteroskedasticity and lower values of α. With small or moderate covariate skew, the homoskedastic saddlepoint test has very accurate rejection rates when the degree of heteroskedasticity is low—that is, when the working model under which it is derived is not too discrepant from the true data-generating model. Just as with the empirical version, the rejection rates of the homoskedastic saddlepoint exceed the nominal level under high skewness, and to an extent that grows with the degree of heteroskedasticity.

In contrast to the saddlepoint tests, the empirical Satterthwaite test has above-nominal rejection rates under most conditions. Like the homoskedastic saddlepoint, the rejection rates of the homoskedastic Satterthwaite test are very accurate when skewness is small or moderate. With moderate or high skewness and strong heteroskedasticity, its rejection rates remain even closer to nominal than those of the homoskedastic saddlepoint test. Of the four tests depicted in this figure, we focus further analysis on the homoskedastic saddlepoint and Satterthwaite approximations.

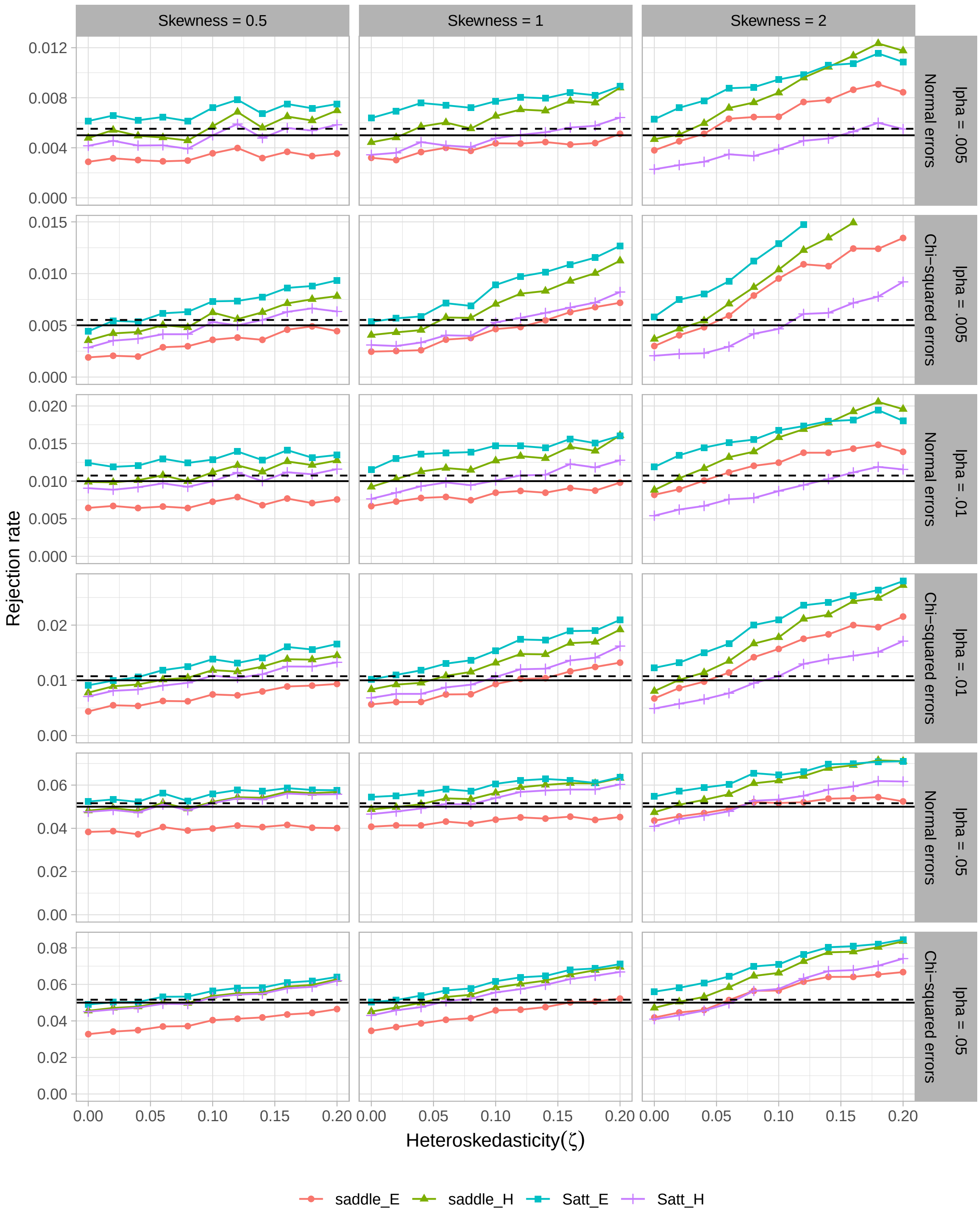

Figure 3. Rejection rates of Satterthwaite and saddlepoint approximation tests for $n = 50$. The solid horizontal line indicates the stated α level and the dashed line indicates an upper confidence bound on simulation error.

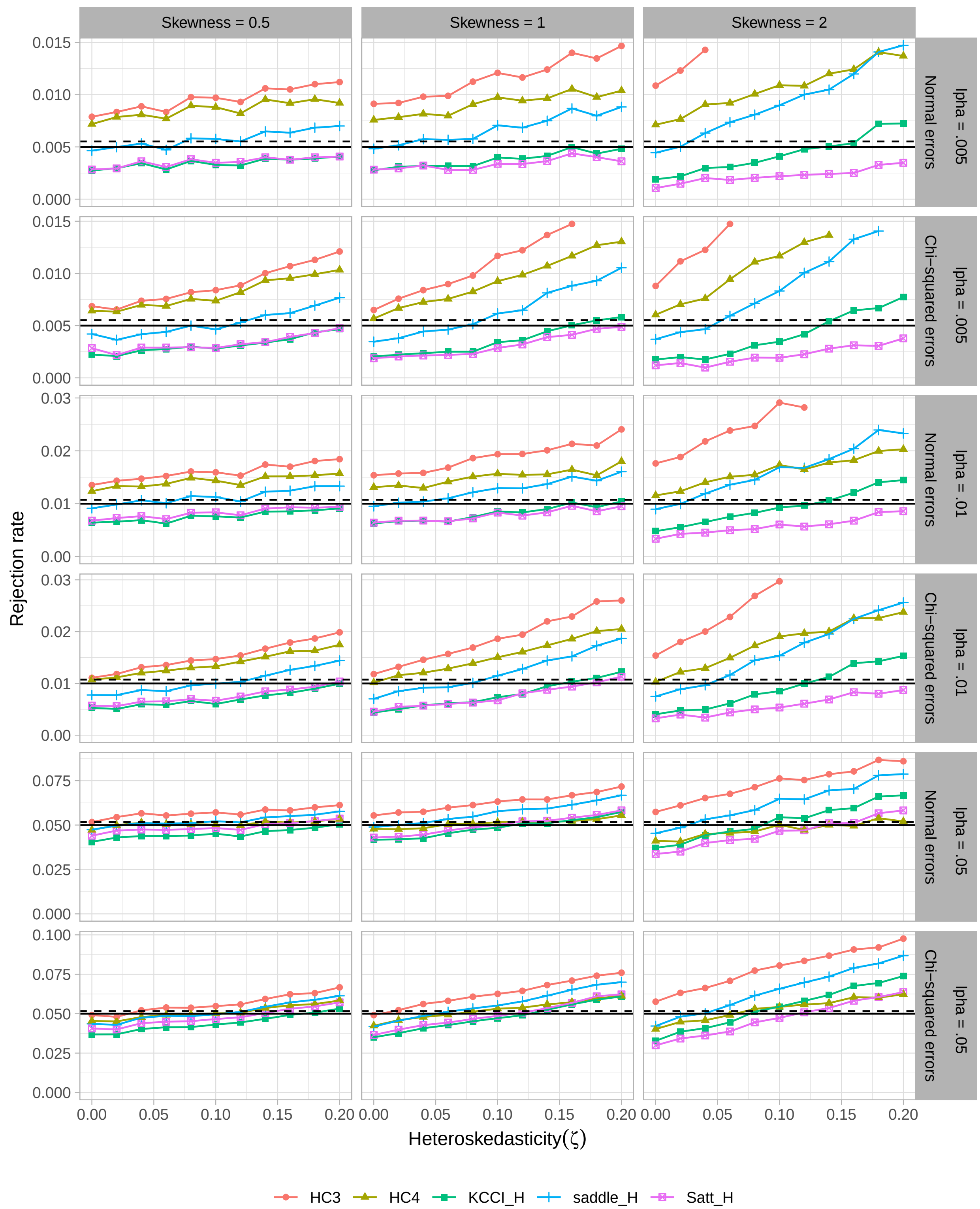

Figure 4. Rejection rates of selected tests for $n = 25$. The solid horizontal line indicates the stated α level and the dashed line indicates an upper confidence bound on simulation error.

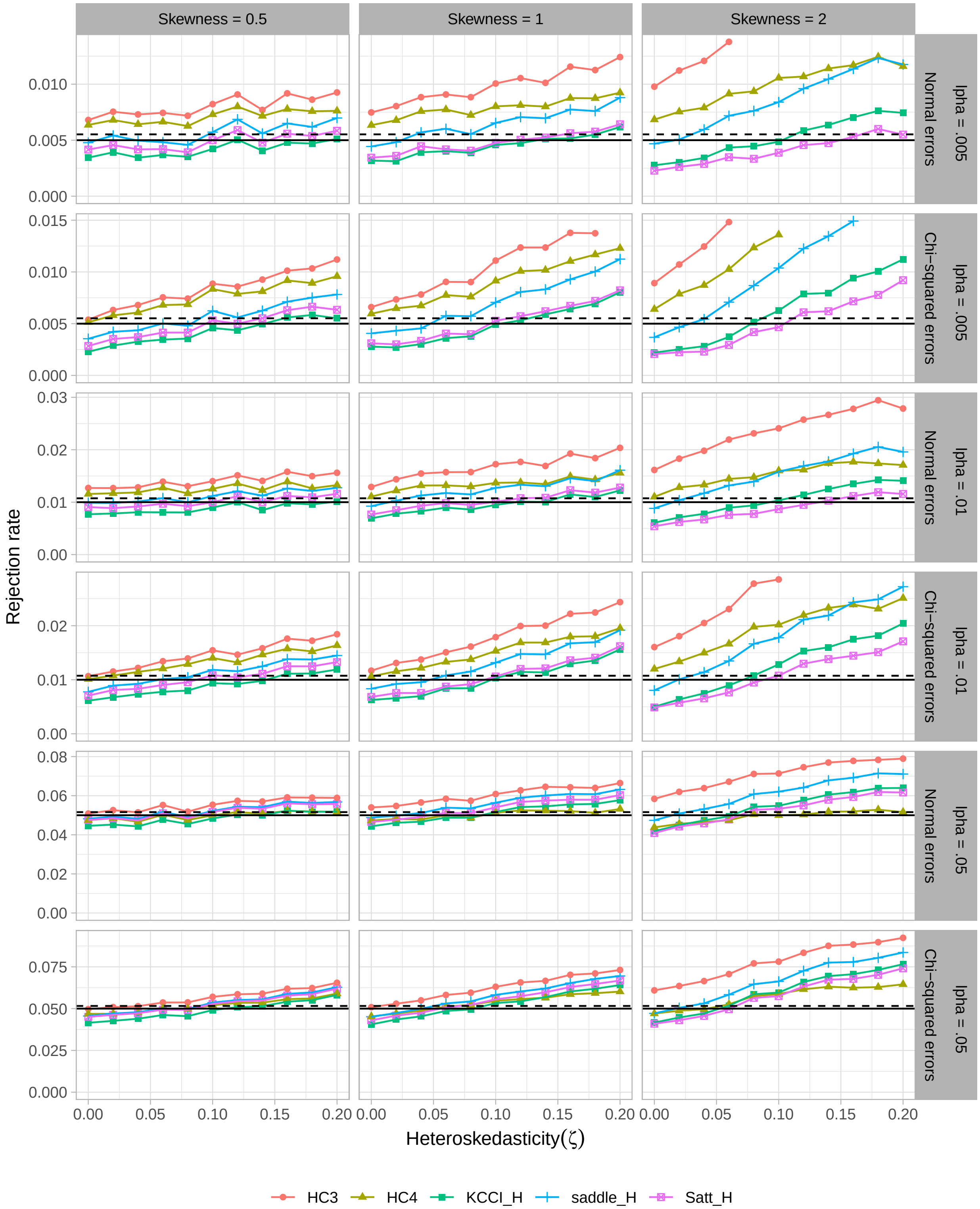

Figure 5. Rejection rates of selected tests for $n = 50$. The solid horizontal line indicates the stated α level and the dashed line indicates an upper confidence bound on simulation error.

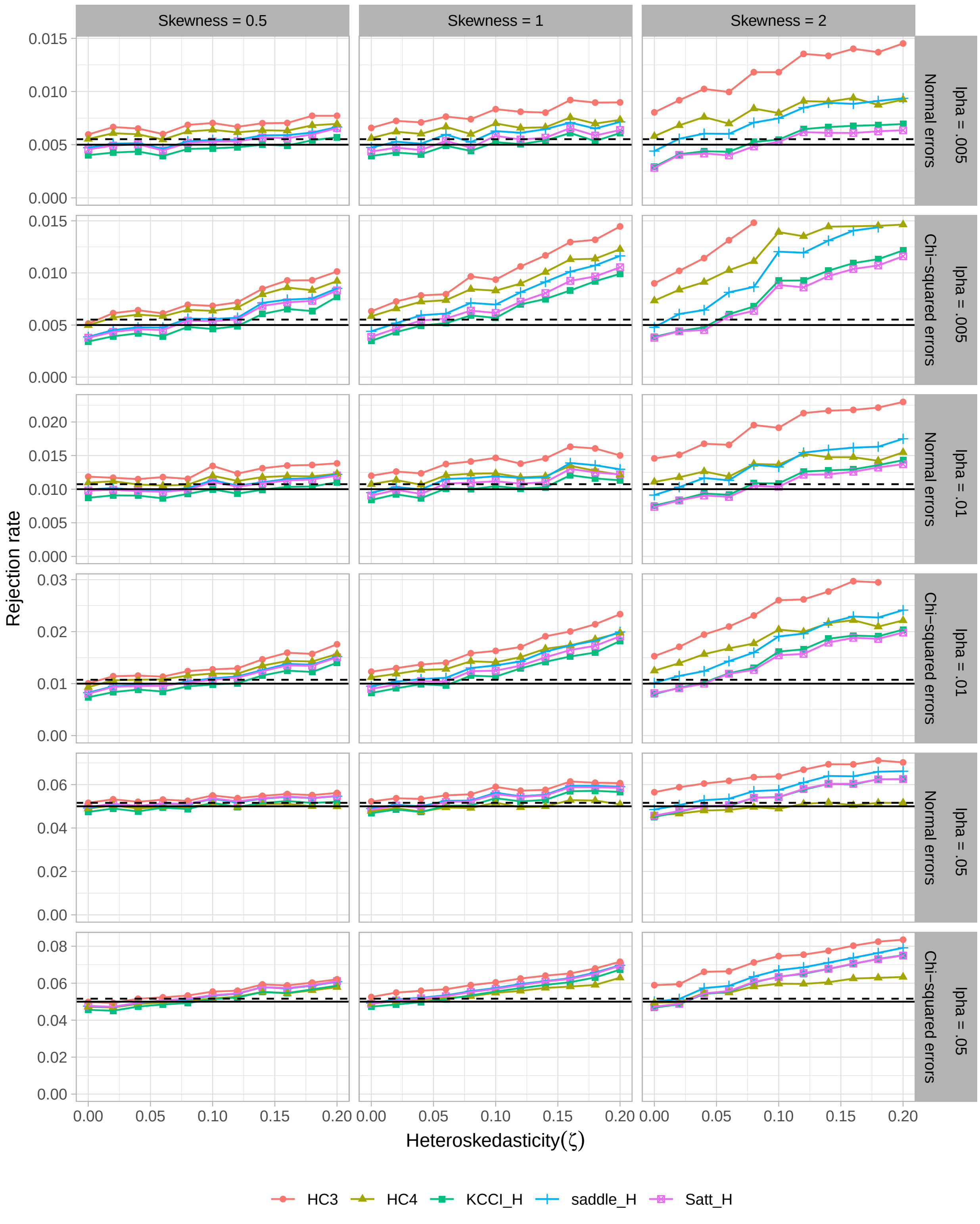

Figure 6. Rejection rates of selected tests for $n = 100$. The solid horizontal line indicates the stated α level and the dashed line indicates an upper confidence bound on simulation error.

Figures 4, 5, and 6 depict the rejection rates of selected tests at sample sizes of $n = 25$, 50, and 100, respectively, so as to allow for direct comparisons. Three trends are apparent. First, the most widely known test, which uses conventional t critical values and HC3, has above-nominal rejection rates under nearly all conditions. The other tests, including the conventional test with HC4 and the homoskedastic versions of Kauermann and Carroll's Edgeworth approximation, the saddlepoint approximation, and the Satterthwaite approximation, all have more accurate rejection rates across the conditions and sample sizes considered.

Second, the conventional test with HC4 has the most accurate rejection rates at the nominal $\alpha = .05$ level. However, for $\alpha = .005$ and $\alpha = .01$, the conventional HC4 test rejects at above-nominal levels. For these smaller values of α, the Edgeworth, saddlepoint, and Satterthwaite tests are closer to (or below) nominal levels across nearly all conditions.

Third, the rejection rates of the three distributional approximation tests are largely very similar when the covariate is mildly or moderately skewed, although some differences appear when the covariate is strongly skewed. In particular, the Edgeworth and Satterthwaite tests maintain rejection rates that are closer to the nominal level than those of the saddlepoint test when the covariate is strongly skewed, and as the degree of heteroskedasticity increases.

Overall, Kauermann and Carroll's Edgeworth approximation and the Satterthwaite approximation (both based on a homoskedastic working model) provide more accurate rejection rates than the conventional test with HC4 for nominal α levels of .005 and .01. For $\alpha = .05$, the conventional test with HC4 provides the most accurate rejection rates, although the Edgeworth and Satterthwaite approximations are quite close when the covariate has minor or moderate skewness.

Discussion

Since Long and Ervin (2000) demonstrated its superiority over the then-current standard HC0 estimator, the conventional t-test with HC3 has become likely the most widely known small-sample correction for hypothesis tests based on HCCMEs. However, results of our simulation exercises demonstrate that this test has inadequate size properties, often rejecting the null more often than the nominal α. At the $\alpha = .05$ level, the conventional test with HC4 maintains the most accurate rejection rates of all of the tests considered here. However, at the $\alpha = .01$ and $\alpha = .005$ levels, several less widely studied tests, which involve different approximation techniques for the reference distribution of the t statistic, have better size properties than even the best-performing conventional test based on HC4. These smaller α levels are relevant in applied work in many situations. For example, in an analysis that uses a multiple comparisons technique to control the family-wise error rate at the .05 level, the per-test α will necessarily be lower than .05.

We must emphasize that the extent to which the findings generalize to other conditions remains unknown. The simulation exercises were limited to a very simple data-generating model with a single covariate and a certain skedasticity function. Moreover, we examined only a limited set of covariate distributions, of varying degrees of skewness. Further investigation is required to understand the relative and absolute performance of these tests under a broader array of data-generating processes.

Our review of HCCME-based hypothesis-testing procedures was also limited in scope. Whereas we considered the case where the design matrix has fixed dimension, recent work by Cattaneo, Jansson, and Newey (2018) proposed HCCMEs that are asymptotically valid when the number of predictors grows with sample size. Another important class of techniques for approximating the distribution of test statistics based on HCCMEs, not considered here, is via bootstrap resampling. Recent attention has focused on a wild bootstrap technique proposed by Liu (1988), which is valid under heteroskedasticity and provides substantially more accurate rejection rates than standard approaches in small samples (Davidson & Flachaire, 2008; Flachaire, 2005). There are several nuances involved in implementing accurate wild bootstrap tests, including how to adjust the residuals, the choice of auxilliary distributions, and whether to bootstrap under a restricted model

(MacKinnon, 2013). In further work, we plan to investigate the performance of the best-performing methods identified in this paper compared to resampling tests such as wild bootstrapping and other recently proposed re-sampling methods (e.g. Richard, 2016).

This review was also limited to methods for testing a single linear contrast. A small amount of work exists on methods for joint tests of multiple parameter constraints in linear regression. Cai and Hayes (2008) proposed a test for joint hypotheses in linear regressions with HCCMEs, which directly extends the empirical Satterthwaite approximation of Lipsitz et al. (1999). Zhang (2012a; 2012b; 2013) proposed a different generalization of the empirical Satterthwaite approximation for univariate and multivariate analysis of variance. Tipton and Pustejovsky (2015) examined the performance of these tests, as well as several novel variations, in the context of meta-regression models with dependent effect sizes. It would be useful to further evaluate the performance of these methods for the simpler case of linear regression models.

Recent work on robust testing has focused largely on adjustments to the HCCMEs themselves (e.g., Hartigan, 2016; Li, Zhang, Zhang, & Wang, 2016) rather than to the reference distribution. Despite the limitations we have noted, we would argue that the distributional approximations that we have reviewed—including Satterthwaite, saddlepoint, and Edgeworth approximation—warrant greater consideration for applications and further attention in methodological research on robust inference. A further reason for focusing on reference distribution approximations is that they are extensible to more general forms of sandwich estimation, whereas the HC4 and HC5 adjustments do not have such clear extensions. The Satterthwaite, saddlepoint, and Edgeworth approximation can all readily be extended to more general models, such as linear regression models estimated by weighted least squares. In fact, McCaffrey and Bell (2006) developed the Satterthwaite and saddlepoint approximations under the even more general framework of generalized estimating equations and cluster-robust covariance matrix estimators; the Kauermann and Carroll (2001) Edgeworth approximation can also be applied directly in this case.

References

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, *49*(4), 431–434.

Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, *28*(2), 169–181.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... others (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6.

Cai, L., & Hayes, A. F. (2008). A new test of linear hypotheses in OLS regression under heteroscedasticity of unknown form. *Journal of Educational and Behavioral Statistics*, *33*(1), 21–40. doi: 10.3102/1076998607302628

Cattaneo, M. D., Jansson, M., & Newey, W. K. (2018). Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association*, *113*(523), 1350–1361.

Chesher, A., & Austin, G. (1991). The finite-sample distributions of heteroskedasticity robust Wald statistics. *Journal of Econometrics*, *47*(1), 153–173. doi: 10.1016/0304-4076(91)90082-O

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics and Data Analysis*, *45*(2), 215–233. doi: 10.1016/S0167-9473(02)00366-3

Cribari-Neto, F., & da Silva, W. B. (2011). A new heteroskedasticity-consistent covariance matrix estimator for the linear regression model. *Advances in Statistical Analysis*, *95*(2), 129–146. doi: 10.1007/s10182-010-0141-2

Cribari-Neto, F., Souza, T. C., & Vasconcellos, K. L. P. (2007). Inference under heteroskedasticity and leveraged data. *Communications in Statistics - Theory and Methods*, *36*(10), 1877–1888. doi: 10.1080/03610920601126589

Davidson, R., & Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of*

Econometrics, *146*(1), 162–169. doi: 10.1016/j.jeconom.2008.08.003

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics.* New York, NY: Oxford University Press.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (pp. 59–82). Berkeley, CA: University of California Press.

Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: Wild bootstrap vs. pairs bootstrap. *Computational Statistics and Data Analysis*, *49*(2), 361–376. doi: 10.1016/j.csda.2004.05.018

Fuller, W. A. (1975). Regression analysis for sample survey. *Sankhya Series C*, *37*, 117–132.

Goutis, C., & Casella, G. (1999). Explaining the saddlepoint approximation. *The American Statistician*, *53*(3), 216–224. doi: 10.1080/00031305.1999.10474463

Hartigan, L. (2016). *Alternative HAC Covariance Matrix Estimators with Improved Finite Sample Properties.* Retrieved from http://ssrn.com/abstract=2783612

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (pp. 221–233). Berkeley, CA: University of California Press.

Huzurbazar, S. (1999). Practical saddlepoint approximations. *The American Statistician*, *53*(3), 225–232. doi: 10.1080/00031305.1999.10474464

Imbens, G. W., & Kolesar, M. (2015). *Robust standard errors in small samples: Some practical advice.* Retrieved from https://www.princeton.edu/{~}mkolesar/papers/small-robust.pdf

Kauermann, G., & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, *96*(456), 1387–1396.

Li, S., Zhang, N., Zhang, X., & Wang, G. (2016). A new heteroskedasticity-consistent covariance matrix estimator and inference under heteroskedasticity. *Journal of Statistical Computation and Simulation*, *87*(1), 198–210. doi: 10.1080/00949655.2016.1198906

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Lipsitz, S. R., Ibrahim, J. G., & Parzen, M. (1999). A degrees-of-freedom approximation for a t-statistic with heterogeneous variance. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *48*(4), 495–506. doi: 10.1111/1467-9884.00207

Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *The Annals of Statistics*, *16*(4), 1696–1708. doi: 10.1214/aos/1176351062

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), 217–224. doi: 10.1080/00031305.2000.10474549

Lugannani, R., & Rice, S. (1980, jun). Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability*, *12*(2), 475. doi: 10.2307/1426607

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In X. Chen & N. R. Swanson (Eds.), *Recent advances and future directions in causality, prediction, and specification analysis.* New York, NY: Springer New York. doi: 10.1007/978-1-4614-1653-1

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

Mathai, A. M., & Provost, S. B. (1992). *Quadratic forms in random variables: theory and applications.* M. Dekker, New York.

McCaffrey, D. F., & Bell, R. M. (2006). Improved hypothesis testing for coefficients in generalized estimating equations with small samples of clusters. *Statistics in* *medicine*, *25*(23), 4081–98. doi: 10.1002/sim.2502

Richard, P. (2016). Heteroskedasticity-robust tests with minimum size distortion. *Communications in Statistics - Theory and Methods*, *0926*, 1–15. doi: 10.1080/03610926.2015.1125497

Rothenberg, T. (1988). Approximate power functions for some robust tests of regression coefficients. *Econometrica*, *56*(5), 997–1019.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, *2*(6), 110–114.

Skinner, C. J. (1989). Domain means, regression and multivariate analyses. In C. J. Skinner, D. Holt, & T. F. Smith (Eds.), *Analysis of complex surveys* (pp. 59–88). New York, NY: John Wiley & Sons.

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, *40*(6), 604–634.

Welch, B. (1947). The generalization of Student's' problem when several different population variances are involved. *Biometrika*, *34*(1/2), 28–35.

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

White, H. (1984). *Asymptotic theory for econometricians.* Orlando, FL: Academic Press, Inc.

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, *11*(10), 1–17.

Zhang, J.-T. (2012a). An approximate degrees of freedom test for heteroscedastic two-way ANOVA. *Journal of Statistical Planning and Inference*, *142*(1), 336–346. doi: 10.1016/j.jspi.2011.07.023

Zhang, J.-T. (2012b). An approximate Hotelling T2 -test for heteroscedastic one-way MANOVA. *Open Journal of Statistics*, *2*, 1–11.

Zhang, J.-T. (2013). Tests of linear hypotheses in the ANOVA under heteroscedasticity. *International Journal of Advanced Statistics and Probability*, *1*(2), 9–24.