Econometric Theory, **41**, 2025, 249–301.
doi:10.1017/S0266466623000269

VALID HETEROSKEDASTICITY ROBUST TESTING

Benedikt M. Pötscher
University of Vienna

David Preinerstorfer
University of St. Gallen

Tests based on heteroskedasticity robust standard errors are an important technique in econometric practice. Choosing the right critical value, however, is not simple at all: conventional critical values based on asymptotics often lead to severe size distortions, and so do existing adjustments including the bootstrap. To avoid these issues, we suggest to use smallest size-controlling critical values, the generic existence of which we prove in this article for the commonly used test statistics. Furthermore, sufficient and often also necessary conditions for their existence are given that are easy to check. Granted their existence, these critical values are the canonical choice: larger critical values result in unnecessary power loss, whereas smaller critical values lead to overrejections under the null hypothesis, make spurious discoveries more likely, and thus are invalid. We suggest algorithms to numerically determine the proposed critical values and provide implementations in accompanying software. Finally, we numerically study the behavior of the proposed testing procedures, including their power properties.

1. INTRODUCTION

Testing hypotheses on the parameters in a regression model with potentially heteroskedastic errors is an important problem in econometrics and statistics (see MacKinnon, 2013 for a recent survey). Since the classical t-statistic (F-statistic, respectively) is not pivotal, or asymptotically pivotal, in such a case in general, even under Gaussianity of the errors, so-called heteroskedasticity robust (aka heteroskedasticity consistent) modifications of these test statistics have been proposed, which are asymptotically standard normally (chi-square, respectively) distributed under the null. These modifications date back to Eicker (1963, 1967) (see also Hinkley, 1977) and have later been popularized in econometrics by White (1980) with great success (see MacKinnon, 2013). Unfortunately, it turned out that tests obtained from these heteroskedasticity robust test statistics by relying on critical values derived from the respective asymptotic null distributions have a

Financial support of the second author by the Program of Concerted Research Actions (ARC) of the Université libre de Bruxelles in an early stage of this project is gratefully acknowledged. We thank two referees, a Co-Editor, and the Editor for helpful comments. Address correspondence to Benedikt Pötscher, Department of Statistics, University of Vienna, A-1090 Oskar-Morgenstern Platz 1, Vienna, Austria; e-mail: benedikt.poetscher@univie.ac.at.

tendency to overreject the null hypothesis in finite samples (and thus are invalid), especially so if the design matrix contains high-leverage points (see, e.g., Davidson and MacKinnon, 1985; MacKinnon and White, 1985; Chesher and Jewitt, 1987). One factor contributing to this overrejection tendency is a downward bias in the covariance matrix estimators used in these test statistics (see Chesher and Jewitt, 1987). In an attempt to reduce the overrejection problem, variants of the before-mentioned heteroskedasticity robust test statistics (often denoted by HC1 through HC4, with HC0 denoting the original proposal) have been considered (see Hinkley, 1977; MacKinnon and White, 1985; Cribari-Neto, 2004).[1] These variants rescale the least-squares residuals before computing the covariance matrix estimator employed in the construction of the test statistic. According to simulation studies reported in, e.g., Davidson and MacKinnon (1985) and Cribari-Neto (2004), these modifications, especially HC3 and HC4, seem to ameliorate the overrejection problem to some extent, but do not eliminate it. Further numerical results are provided in Chesher and Austin (1991) (see also Chesher, 1989). Numerical results in Section 11 confirm these observations. Variants of HC0–HC3, denoted by HC0R–HC3R, obtained by using restricted instead of unrestricted least-squares residuals in the computation of the covariance matrix estimators employed by the various test statistics (the restriction alluded to being the restriction defining the null hypothesis), have been introduced in Davidson and MacKinnon (1985). In their simulation experiments, this typically leads to tests that do not overreject, but that may substantially underreject; see also the simulation results in Godfrey (2006), who additionally also considers HC4R. However, as will be shown in Section 11, also these tests are in general not immune to (sometimes substantial) overrejection.

Note that, under the typical assumptions used in the literature, all the modifications of HC0 discussed so far have the same asymptotic distribution as HC0, and thus the same critical value as for HC0 (obtained from the asymptotic null distribution) is also used for these modifications in the before-mentioned literature. Sometimes small-sample adjustments to the asymptotic critical values are attempted by using the quantiles from a t_d-distribution rather than from the asymptotic normal distribution, where the degrees of freedom d are either set to $n-k$ (n and k denoting sample size and number of regressors, respectively), or are obtained through proposals set down by Satterthwaite (1946) or Bell and McCaffrey (2002) (see also Imbens and Kolesár, 2016). While these adjustments can lead to improvements, numerical results presented in Section 11 show that these adjustments are also not able to solve the overrejection problem in general. An alternative approach is to use bootstrap methods to compute critical values for the test statistics HC0–HC4 or HC0R–HC4R. The relevant literature is reviewed in Pötscher and Preinerstorfer (2023), and it is shown that such methods are again not immune to the overrejection problem in general.[2] A referee has pointed out

[1] For a recent contribution geared toward high-dimensional models, see Cattaneo, Jansson, and Newey (2018).

[2] Another possibility is to use Edgeworth expansions to find better critical values (see Rothenberg, 1988 for the case of the HC0 test statistic and Davidson and MacKinnon, 1985 for the HC0R test statistic). Simulation results in Davidson

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

the recent papers by Chu et al. (2021) and Hansen (2021), both of which propose a testing procedure that can be viewed as a parametric bootstrap method.[3] No theoretical justification is given in those papers. In fact, as we show in Appendix G of the Supplementary Material, the proposed procedures can be considerably oversized, a feature that can already be seen to some extent in the numerical results given in Chu et al. (2021) and Hansen (2021).

A result by Bakirov and Székely (2005) needs to be mentioned here which states that—in the special case of testing a hypothesis on the location parameter of a heteroskedastic location model with errors that are Gaussian or scale mixtures thereof—the classical two-sided *t*-test (with the usual critical value) has null rejection probability not exceeding the nominal significance level under *any* form of heteroskedasticity (for a certain range of significance levels); see Ibragimov and Müller (2010) for more discussion. Ibragimov and Müller (2016), extending a result in Mickey and Brown (1966), provide a related result in the case of the comparison of two heteroskedastic populations (see also Bakirov, 1998). Section 5.2 provides some more discussion. We note that all results mentioned in that section are applicable only to testing *certain scalar* linear contrasts.

Except for the Bakirov and Székely (2005) result and the variations discussed in Section 5.2, which apply only to quite special situations like, e.g., the heteroskedastic location model, none of the methods discussed so far comes with a theoretical result implying that their associated (finite sample) null rejection probabilities are guaranteed not to exceed the nominal significance level whatever the form of heteroskedasticity may be. [4] In fact, it transpires from the preceding discussion and the numerical results in Section 11 that for *any* of these methods, instances of testing problems can be found for which the method in question overrejects substantially. Therefore, it is imperative to be able to find size-controlling critical values for the test statistics considered, i.e., critical values such that the resulting worst-case rejection probability under the null hypothesis does not exceed the nominal significance level. We shall, hence, pursue in this article the construction of size-controlling critical values for the test statistics HC0–HC4, HC0R–HC4R, as well as for (two variants of) the classical (i.e., uncorrected) *F*-statistic (including the absolute value of the *t* -statistic as a special case).

In the present article, we consider classes of test statistics that contain the beforementioned heteroskedasticity robust test statistics as special cases and show under which conditions—and how—a critical value can be found such that the resulting

and MacKinnon (1985) and MacKinnon and White (1985) indicate that this does not work too well in practice. Of course, such expansions could also be worked out for the other versions of the test statistics mentioned, but this does not seem to have been pursued in the literature.

[3] A reader has pointed out that the results in Phillips (1993) could also be developed into numerical approximations similar to the ones in Hansen (2021).

[4] In the special case where the number of restrictions tested equals the number of regression parameters, Davidson and Flachaire (2008) have a result which implies that certain wild bootstrap-based heteroskedasticity robust tests have size equal to the nominal significance level (and hence do not overreject) in finite samples. We note that this result in Davidson and Flachaire (2008) is not entirely correct as stated, but needs some amendments and corrections (see Pötscher and Preinerstorfer, 2023).

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

test is guaranteed to have size less than or equal to α, the prescribed significance level.[5] It turns out that the conditions for size controllability are broadly satisfied; in particular, for the commonly used test statistics, they are satisfied *generically* in a sense made precise further below.

We want to emphasize that the existence of size-controlling critical values for heteroskedasticity robust test statistics is *not* a trivial matter, as it has been shown in Preinerstorfer and Pötscher (2016, Sect. 4) that there are cases where the size of such tests is always one, *regardless* of the choice of critical value (see also the discussion in Proposition 5.7 further below). And, even in cases where size control is possible by an appropriate choice of critical value, the standard critical values proposed in the literature (including the small-sample adjustments discussed above) are *not* guaranteed to deliver size control; in fact, they may fail to do so by a considerable margin (i.e., they are much too small to control size at the desired level) as shown in Section 11. Our theoretical results also show the existence of a computable "threshold" C^*, say, such that any critical value C satisfying $C < C^*$ necessarily leads to a test with size 1 (see Proposition 5.5). Since C^* is not difficult to compute, it can be used as a simple check to weed out unsuitable proposals for critical values.

Apart from avoiding overrejection by construction, the use of smallest size-controlling, rather than conventional, critical values offer also advantages in terms of power in instances where conventional critical values lead to underrejection (i.e., lead to a worst-case rejection probability under the null hypothesis less than the nominal significance level) as is sometimes the case (see Sections 6.2.2 and 11.2). In fact, once one has decided on a test statistic to be used for the given null hypothesis, using the smallest size-controlling critical value (provided it exists) is obviously the optimal way to proceed.

We also discuss how the critical values that lead to size control can be determined numerically and provide the R-package **hrt** (Preinerstorfer, 2021) for their computation. The usefulness of the proposed algorithms and their implementation in the R-package are illustrated numerically on some testing problems in Section 11. In particular, we compare tests obtained from various of the abovementioned test statistics when used with smallest size-controlling critical values in terms of their power functions. The package **hrt** also contains a routine for determining the size of a test obtained from a user-supplied critical value. It is important to note that if in a particular application one uses the observed value of the test statistic as the user-supplied critical value in this routine, this routine actually returns a "valid p-value" in the following sense: checking whether or not this "p-value" is smaller than the prescribed significance level α is equivalent to checking whether or not the observed value of the test statistic is larger than or equal to the smallest

[5] A less principled attempt at finding a valid test in a *given* testing problem (i.e., for given design matrix and restriction to be tested) could consist in the practitioner studying the size of a handful of tests (obtained from a few of the abovementioned test statistics in conjunction with a few of the proposed critical values) by means of an extensive Monte Carlo study and in hoping that one of the test procedures emerges from this study as valid for the particular testing problem at hand. Besides being a numerically costly procedure, it does not come with any guarantee of success.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

size-controlling critical value. Note that the former check avoids the need to actually compute the smallest size-controlling critical value, which is advantageous from a computational point of view. See Section 10 for more details.

In the article, we work under a Gaussianity assumption. We stress, however, that this assumption is mainly made for convenience of presentation; as shown in Section 7.1, this assumption can be relaxed considerably.

While a trivial remark, we would like to note that the size-control results given in this article can easily be translated into results stating that the minimal coverage probability of the associated confidence set obtained by "inverting" the test is not less than the nominal confidence level.

The article is organized as follows: after introducing notation and the most important test statistics in Sections 2 and 3, Section 4 provides some intuition for our size-control results which are presented in Sections 5 and 6, with some further results relegated to Appendix A of the Supplementary Material. Section 7 discusses ways of relaxing the underlying assumptions. Possible extensions to other classes of test statistics are discussed in Section 8, whereas a few comments on power are collected in Section 9. Section 11 provides the numerical results including a power study, with some details relegated to Appendix F of the Supplementary Material. Section 12 concludes. Proofs and some technical results can be found in Appendixes B–D of the Supplementary Material. The algorithms for computing rejection probabilities (including size) and smallest size-controlling critical values are outlined in Section 10, and are presented in detail in Appendix E of the Supplementary Material. Appendix G of the Supplementary Material contains a discussion of Chu et al. (2021) and Hansen (2021).

2. FRAMEWORK

Consider the linear regression model

$$\mathbf{Y} = X\beta + \mathbf{U}, \tag{1}$$

where X is a (real) nonstochastic regressor (design) matrix of dimension $n \times k$ and where $\beta \in \mathbb{R}^k$ denotes the unknown regression parameter vector. We always assume $\operatorname{rank}(X) = k$ and $1 \leq k < n$. We furthermore assume that the $n \times 1$ disturbance vector $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)'$ has mean zero and unknown covariance matrix $\sigma^2\Sigma$, where Σ varies in a user-specified (nonempty) set $\mathfrak{C}$ describing the allowed forms of heteroskedasticity, with $\mathfrak{C}$ satisfying $\mathfrak{C} \subseteq \mathfrak{C}_{Het}$, and where $0 < \sigma^2 < \infty$ holds (σ always denoting the positive square root).[6] The set $\mathfrak{C}$ will be referred to as the "heteroskedasticity model." Here,

$$\mathfrak{C}_{Het} = \left\{ \operatorname{diag}(\tau_1^2, \ldots, \tau_n^2) : \tau_i^2 > 0 \text{ for all } i, \sum_{i=1}^{n} \tau_i^2 = 1 \right\},$$

[6] Since we are concerned with finite-sample results only, the elements of $\mathbf{Y}$, X, and $\mathbf{U}$ (and even the probability space supporting $\mathbf{Y}$ and $\mathbf{U}$) may depend on sample size n, but this will not be expressed in the notation. Furthermore, the obvious dependence of $\mathfrak{C}$ on n will also not be shown in the notation.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

where $\operatorname{diag}(\tau_1^2, \dots, \tau_n^2)$ denotes the $n \times n$ matrix with diagonal elements given by τ_i^2. That is, the errors in the regression model are uncorrelated but can be heteroskedastic. In particular, if $\mathfrak{C}$ is chosen to be $\mathfrak{C}_{Het}$, one allows for heteroskedasticity of completely unknown form. The normalization condition $\sum_{i=1}^n \tau_i^2 = 1$ is included here only in order to guarantee identifiability of σ^2 and Σ, and could be replaced by any other normalization condition, such as $\max \tau_i^2 = 1$, or $\tau_1^2 = 1$, without affecting the final results (because any of these normalizations leads to the same overall set of covariance matrices $\sigma^2\Sigma$ when σ^2 varies through the positive real line). Although a trivial observation, we stress the fact that all conceivable forms of heteroskedasticity, including parametric ones, can (possibly after normalization) be cast as submodels $\mathfrak{C}$ of $\mathfrak{C}_{Het}$.

Mainly for ease of exposition, we shall maintain in the sequel that the disturbance vector $\mathbf{U}$ *is normally distributed. This assumption can be substantially relaxed as discussed in Section 7.1.* The linear model described in (1), together with the just made Gaussianity assumption on $\mathbf{U}$ and with the given heteroskedasticity model $\mathfrak{C}$, then induces a collection of distributions on the Borel-sets of $\mathbb{R}^n$, the sample space of $\mathbf{Y}$. Denoting a Gaussian probability measure with mean $\mu \in \mathbb{R}^n$ and (possibly singular) covariance matrix A by $P_{\mu,A}$, the induced collection of distributions is then given by

$$\left\{P_{\mu,\sigma^2\Sigma} : \mu \in \operatorname{span}(X), 0 < \sigma^2 < \infty, \Sigma \in \mathfrak{C}\right\}, \tag{2}$$

where $\operatorname{span}(X)$ denotes the column space of X. Since every $\Sigma \in \mathfrak{C}$ is positive definite by assumption, each element of the set in the previous display is absolutely continuous with respect to (w.r.t.) Lebesgue measure on $\mathbb{R}^n$.

We shall consider the problem of testing a linear (better: affine) hypothesis on the parameter vector $\beta \in \mathbb{R}^k$, i.e., the problem of testing the null $R\beta = r$ against the alternative $R\beta \neq r$, where R is a $q \times k$ matrix always of rank $q \geq 1$ and $r \in \mathbb{R}^q$. Set $\mathfrak{M} = \operatorname{span}(X)$. Define the affine space

$$\mathfrak{M}_0 = \{\mu \in \mathfrak{M} : \mu = X\beta \text{ and } R\beta = r\},$$

and let

$$\mathfrak{M}_1 = \{\mu \in \mathfrak{M} : \mu = X\beta \text{ and } R\beta \neq r\}.$$

Adopting these definitions, this testing problem can then be written more precisely as

$$H_0 : \mu \in \mathfrak{M}_0,\ 0 < \sigma^2 < \infty,\ \Sigma \in \mathfrak{C} \quad \text{vs.} \quad H_1 : \mu \in \mathfrak{M}_1,\ 0 < \sigma^2 < \infty,\ \Sigma \in \mathfrak{C}. \tag{3}$$

With $\mathfrak{M}_0^{lin}$, we shall denote the linear space parallel to $\mathfrak{M}_0$, i.e., $\mathfrak{M}_0^{lin} = \mathfrak{M}_0 - \mu_0 = \{X\beta : R\beta = 0\}$, where $\mu_0 \in \mathfrak{M}_0$. Of course, $\mathfrak{M}_0^{lin}$ does not depend on the choice of $\mu_0 \in \mathfrak{M}_0$.

As already mentioned, the assumption of Gaussianity is made mainly for simplicity of presentation and can be relaxed substantially (see Section 7.1). The assumption of nonstochastic regressors entails little loss of generality either, which

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

is important to emphasize: if X is random and $\mathbf{U}$ is conditionally on X distributed as $N(0,\sigma^2\Sigma)$, with $\sigma^2 = \sigma^2(X) > 0$ and $\Sigma = \Sigma(X) \in \mathfrak{C}_{Het}$, the results of the article can be applied after one conditions on X (and a similar statement applies to the generalizations to non-Gaussianity discussed in Section 7.1). See Section 7.2 for more discussion and details. For arguments supporting conditional inference, see, e.g., Robinson (1979). Note that such a "strict exogeneity" assumption is quite natural in the situation considered here.

We next collect some further terminology and notation used throughout the article. A (nonrandomized) *test* is the indicator function of a Borel-set W in $\mathbb{R}^n$, with W called the corresponding *rejection region*. The *size* of such a test (rejection region) is—as usual—defined as the supremum over all rejection probabilities under the null hypothesis H_0 given in (3), i.e.,

$$\sup_{\mu\in\mathfrak{M}_0}\ \sup_{0<\sigma^2<\infty}\ \sup_{\Sigma\in\mathfrak{C}} P_{\mu,\sigma^2\Sigma}(W).$$

In slight abuse of terminology, we shall sometimes refer to this quantity as "the size of W over $\mathfrak{C}$" when we want to emphasize the rôle of $\mathfrak{C}$. Throughout the article, we let $\hat{\beta}(y) = \left(X'X\right)^{-1}X'y$, where X is the design matrix appearing in (1) and $y \in \mathbb{R}^n$. The corresponding ordinary least-squares (OLS) residual vector is denoted by $\hat{u}(y) = y - X\hat{\beta}(y)$, and its elements are denoted by $\hat{u}_t(y)$. The elements of X are denoted by x_{ti}, whereas $x_{t\cdot}$ and $x_{\cdot i}$ denote the tth row and ith column of X, respectively. For $\mathcal{A}$ an affine subspace of $\mathbb{R}^n$ satisfying $\mathcal{A} \subseteq \operatorname{span}(X)$, let $\tilde{\beta}_{\mathcal{A}}(y)$ denote the restricted least-squares estimator, i.e., $X\tilde{\beta}_{\mathcal{A}}(y)$ solves

$$\min_{z\in\mathcal{A}}(y-z)'(y-z).$$

Lebesgue measure on the Borel-sets of $\mathbb{R}^n$ will be denoted by $\lambda_{\mathbb{R}^n}$, whereas Lebesgue measure on an arbitrary affine subspace $\mathcal{A}$ of $\mathbb{R}^n$ (but viewed as a measure on the Borel-sets of $\mathbb{R}^n$) will be denoted by $\lambda_{\mathcal{A}}$, with zero-dimensional Lebesgue measure being interpreted as point mass. The set of real matrices of dimension $l \times m$ is denoted by $\mathbb{R}^{l\times m}$ (all matrices in the article will be real matrices), and Lebesgue measure on this set equipped with its Borel σ-field is denoted by $\lambda_{\mathbb{R}^{l\times m}}$. Let B' denote the transpose of a matrix $B \in \mathbb{R}^{l\times m}$, and let $\operatorname{span}(B)$ denote the subspace in $\mathbb{R}^l$ spanned by its columns. For a symmetric and nonnegative definite matrix B, we denote the unique symmetric and nonnegative definite square root by $B^{1/2}$. For a linear subspace $\mathcal{L}$ of $\mathbb{R}^n$, we let $\mathcal{L}^{\perp}$ denote its orthogonal complement and we let $\Pi_{\mathcal{L}}$ denote the orthogonal projection onto $\mathcal{L}$. The Euclidean norm is denoted by $\|\cdot\|$, but the same symbol is also used to denote a norm of a matrix. The jth standard basis vector in $\mathbb{R}^n$ is written as $e_j(n)$. Furthermore, we let $\mathbb{N}$ denote the set of all positive integers. A sum (product, respectively) over an empty index set is to be interpreted as 0 (1, respectively). Finally, for $\mathcal{A}$ an affine subspace of $\mathbb{R}^n$, let $G(\mathcal{A})$ denote the group of all affine transformations $y \mapsto \delta(y-a)+a^*$, where $\delta \in \mathbb{R}$, $\delta \neq 0$, and a as well as a^* are elements of $\mathcal{A}$; for more information, see Section 5.1 of Preinerstorfer and Pötscher (2016).

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

3. HETEROSKEDASTICITY ROBUST TEST STATISTICS USING UNRESTRICTED RESIDUALS

We now introduce two test statistics that will feature prominently in the following. Variants thereof that use restricted residuals are discussed in Section 6. For results pertaining to other classes of test statistics, see Section 8. The test statistic we shall consider first is a standard heteroskedasticity robust test statistic frequently encountered in the literature. It is given by

$$T_{Het}(y) = \begin{cases} (R\hat{\beta}(y) - r)'\hat{\Omega}_{Het}^{-1}(y)(R\hat{\beta}(y) - r), & \text{if } \det \hat{\Omega}_{Het}(y) \neq 0, \\ 0, & \text{if } \det \hat{\Omega}_{Het}(y) = 0, \end{cases} \tag{4}$$

where $\hat{\Omega}_{Het} = R\hat{\Psi}_{Het}R'$ and where $\hat{\Psi}_{Het}$ is a heteroskedasticity robust estimator as considered in Eicker (1963, 1967), which later on has found its way into the econometrics literature (e.g., White, 1980). It is of the form

$$\hat{\Psi}_{Het}(y) = (X'X)^{-1}X'\operatorname{diag}\left(d_1\hat{u}_1^2(y), \ldots, d_n\hat{u}_n^2(y)\right)X(X'X)^{-1},$$

where the constants $d_i > 0$ sometimes depend on the design matrix. Typical choices for d_i suggested in the literature are $d_i = 1$, $d_i = n/(n-k)$, $d_i = (1-h_{ii})^{-1}$, or $d_i = (1-h_{ii})^{-2}$, where h_{ii} denotes the ith diagonal element of the projection matrix $X(X'X)^{-1}X'$ (see Long and Ervin, 2000 for an overview). Another suggestion is $d_i = (1-h_{ii})^{-\delta_i}$ for $\delta_i = \min(nh_{ii}/k, 4)$ (see Cribari-Neto, 2004). For the last three choices of d_i just given, we use the convention that we set $d_i = 1$ in case $h_{ii} = 1$. Note that $h_{ii} = 1$ implies $\hat{u}_i(y) = 0$ for every y, and hence it is irrelevant which real value is assigned to d_i in case $h_{ii} = 1$.[7] The five examples for the weights d_i just given correspond to what is often called HC0–HC4 weights in the literature.

In conjunction with the test statistic T_{Het}, we shall consider the following mild assumption, which is Assumption 3 in Preinerstorfer and Pötscher (2016). As discussed further below, this assumption is in a certain sense unavoidable when using T_{Het}. It furthermore also entails that our choice of assigning $T_{Het}(y)$ the value zero in case $\hat{\Omega}_{Het}(y)$ is singular has no import on the probabilistic results of the article (because of Lemma 3.1(c) and absolute continuity of the measures $P_{\mu,\sigma^2\Sigma}$).

Assumption 1. Let $1 \leq i_1 < \cdots < i_s \leq n$ denote all the indices for which $e_{i_j}(n) \in \operatorname{span}(X)$ holds where $e_j(n)$ denotes the jth standard basis vector in $\mathbb{R}^n$. If no such index exists, set $s = 0$. Let $X'(\neg(i_1, \ldots i_s))$ denote the matrix which is obtained from X' by deleting all columns with indices i_j, $1 \leq i_1 < \cdots < i_s \leq n$ (if $s = 0$, no column is deleted). Then $\operatorname{rank}\left(R(X'X)^{-1}X'(\neg(i_1, \ldots, i_s))\right) = q$ holds.

Observe that this assumption only depends on X and R and hence can be checked. Obviously, a simple sufficient condition for Assumption 1 to hold is that $s = 0$ (i.e., that $e_j(n) \notin \operatorname{span}(X)$ for all j), a generically satisfied condition.

[7] In fact, $h_{ii} = 1$ is equivalent to $\hat{u}_i(y) = 0$ for every y, each of which in turn is equivalent to $e_i(n) \in \operatorname{span}(X)$.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

Furthermore, we introduce the matrix

$$\begin{aligned} B(y) &= R(X'X)^{-1}X'\operatorname{diag}\left(\hat{u}_1(y),\ldots,\hat{u}_n(y)\right) \\ &= R(X'X)^{-1}X'\operatorname{diag}\left(e_1'(n)\Pi_{\operatorname{span}(X)^{\perp}}y,\ldots,e_n'(n)\Pi_{\operatorname{span}(X)^{\perp}}y\right). \end{aligned} \tag{5}$$

The facts collected in the subsequent lemma, which is taken from Pötscher and Preinerstorfer (2023) (but see also Lemma 4.1 in Preinerstorfer and Pötscher, 2016 and Lemma 5.18 in Pötscher and Preinerstorfer, 2018), will be used in the sequel.

Lemma 3.1.

(a) $\hat{\Omega}_{Het}(y)$ *is nonnegative definite for every* $y \in \mathbb{R}^n$.

(b) $\hat{\Omega}_{Het}(y)$ *is singular (zero, respectively) if and only if* $\operatorname{rank}(B(y)) < q$ $(B(y) = 0$, *respectively).*

(c) *The set* B *given by* $\{y \in \mathbb{R}^n : \operatorname{rank}(B(y)) < q\}$ *(or in view of (b) equivalently given by* $\{y \in \mathbb{R}^n : \det(\hat{\Omega}_{Het}(y)) = 0\}$*) is either a* $\lambda_{\mathbb{R}^n}$*-null set or the entire sample space* $\mathbb{R}^n$*. The latter occurs if and only if Assumption 1 is violated (in which case, the test based on* T_{Het} *becomes trivial, as then* T_{Het} *is identically zero).*

(d) *Under Assumption 1, the set* B *is a finite union of proper linear subspaces of* $\mathbb{R}^n$*; in case* $q = 1$*,* B *is even a proper linear subspace itself.*[8]

(e) B *is a closed set and contains* $\operatorname{span}(X)$*. Furthermore,* B *is* $G(\mathfrak{M})$*-invariant and, in particular,* $\mathsf{B} + \operatorname{span}(X) = \mathsf{B}$ *holds.*

In light of Part (c) of the lemma, we see that Assumption 1 is a natural and unavoidable condition if one wants to obtain a sensible test from T_{Het}.[9] Furthermore, note that, if $\mathsf{B} = \operatorname{span}(X)$ is true, then Assumption 1 must be satisfied (since $\operatorname{span}(X)$ is a $\lambda_{\mathbb{R}^n}$-null set due to the maintained assumption $k < n$). As shown in Lemma A.3 in Pötscher and Preinerstorfer (2018), for any given restriction matrix R, the relation $\mathsf{B} = \operatorname{span}(X)$ holds generically in various universes of design matrices. For later use, we also mention that under Assumption 1, the test statistic T_{Het} is continuous at every $y \in \mathbb{R}^n \backslash \mathsf{B}$.[10]

Next, we also consider the classical (i.e., uncorrected) F-test statistic, i.e.,

$$T_{uc}(y) = \begin{cases} (R\hat{\beta}(y) - r)'\left(\hat{\sigma}^2(y)R\left(X'X\right)^{-1}R'\right)^{-1}(R\hat{\beta}(y) - r), & \text{if } y \notin \operatorname{span}(X), \\ 0, & \text{if } y \in \operatorname{span}(X), \end{cases} \tag{6}$$

where $\hat{\sigma}^2(y) = \hat{u}(y)'\hat{u}(y)/(n-k) \geq 0$ (which vanishes if and only if $y \in \operatorname{span}(X)$). Our choice to set $T_{uc}(y) = 0$ for $y \in \operatorname{span}(X)$ again has no import on the probabilistic results in the article, since $\operatorname{span}(X)$ is a $\lambda_{\mathbb{R}^n}$-null set as a consequence of the maintained assumption that $k < n$ (and since the measures $P_{\mu,\sigma^2\Sigma}$ are absolutely

[8] If Assumption 1 is violated, B equals $\mathbb{R}^n$ by Part (c).

[9] If this assumption is violated, then T_{Het} is identically zero, an uninteresting trivial case.

[10] If Assumption 1 is violated, then T_{Het} is constant equal to zero, and hence is trivially continuous everywhere.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

continuous). For reasons of comparability with (4), we have chosen not to normalize the numerator in (6) by q, the number of restrictions to be tested, as is often done in the definition of the classical F-test statistic. This also has no import on the results as the factor $\frac{1}{q}$ can be absorbed into the critical value. For later use, we also mention that the test statistic T_{uc} is continuous at every $y \in \mathbb{R}^n \backslash \operatorname{span}(X)$.

Remark 3.2.

(i) The test statistics T_{Het} as well as T_{uc} are $G(\mathfrak{M}_0)$-invariant as is easily seen (with the respective exceptional sets B and $\operatorname{span}(X)$ being $G(\mathfrak{M})$-invariant).

(ii) Both statistics actually belong to the class of nonsphericity-corrected F-type test statistics in the sense of Section 5.4 in Preinerstorfer and Pötscher (2016) (terminology being somewhat unfortunate in case of T_{uc} as no correction for the non-sphericity is applied in this case). See Remark C.1 in Appendix C of the Supplementary Material for more discussion.

Remark 3.3. For later use, we note the following: suppose (R,r) and $(\bar{R},\bar{r})$ are both of dimension $q \times (k+1)$ and have $\operatorname{rank}(R) = \operatorname{rank}(\bar{R}) = q$. (i) Then (R,r) and $(\bar{R},\bar{r})$ give rise to the same set $\mathfrak{M}_0$, and thus to the same testing problem (3), if and only if $(AR, Ar) = (\bar{R},\bar{r})$ holds for a nonsingular $q \times q$ matrix A. (ii) The test statistics T_{Het} and T_{uc} remain the same whether they are computed using (R,r) or $(\bar{R},\bar{r})$ provided $(AR, Ar) = (\bar{R},\bar{r})$ holds for a nonsingular $q \times q$ matrix A. (To see this, note that the respective exceptional sets B and $\operatorname{span}(X)$ are the same irrespective of whether (R,r) or $(\bar{R},\bar{r})$ is used, and that A cancels out in the respective quadratic forms appearing in the definitions of the test statistics.)

4. SOME INTUITION ON WHY CONVENTIONAL CRITICAL VALUES CAN LEAD TO OVERREJECTION

We begin the heuristic discussion by considering the testing problem (3) with heteroskedasticity model $\mathfrak{C} = \mathfrak{C}_{Het}$ (i.e., heteroskedasticity of unknown form). Let T stand for any of the test statistics introduced in Section 3, with rejection occurring whenever $T \geq C$, C a critical value.[11] ,[12] For simplicity of presentation, we assume $r = 0$. As discussed in Section 1, basing the test on the conventional critical value $C_{\chi^2(q),0.05}$ (the 95% quantile of a chi-square distribution with q degrees of freedom) often leads to substantial overrejection, i.e., the size of the test (over $\mathfrak{C}_{Het}$) is substantially larger than the desired value $\alpha = 0.05$. One mechanism leading to such overrejection is constituted by a concentration phenomenon discussed at some length in Preinerstorfer and Pötscher (2016): in the present situation, the distribution $P_{0,\sigma^2\Sigma}$ "concentrates" on a so-called concentration subspace (given by $\operatorname{span}(e_i(n))$) when Σ is "close" to one of the singular matrices $e_i(n)e_i(n)'$.[13] In such a case, depending on the design matrix X and the hypothesis given by (R,r),

[11] In case of $T = T_{Het}$, Assumption 1 is supposed to hold.

[12] The discussion similarly applies to the test statistics introduced in Section 6.

[13] There are also other concentration subspaces in the present situation which we can ignore for the heuristic discussion.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

the concentration space may fall into the rejection region $\{T \geq C_{\chi^2(q),0.05}\}$, leading to a rejection probability close to one, and thus much larger than $\alpha = 0.05$.[14] Even if the concentration subspace $\operatorname{span}(e_i(n))$ is not contained in the rejection region, but is sufficiently close to it, a considerable portion of the mass of $P_{0,\sigma^2\Sigma}$ may nevertheless fall into the rejection region if Σ is close to, but not too close to $e_i(n)e_i(n)'$. This again leads to a relatively large rejection probability. Overrejection will often be especially pronounced if certain high-leverage points are present in the design matrix.[15]

In order to obtain a test that has size controlled by α (i.e., size $\leq \alpha$) in situations as just described, the rejection region $\{T \geq C_{\chi^2(q),0.05}\}$ has to be narrowed down, i.e., $C_{\chi^2(q),0.05}$ has to be replaced by a suitably larger critical value C. Whether or not this can successfully be accomplished by a (finite) C, is a nontrivial question, the answer depending on whether or not all possible concentration subspaces can be made to fall outside of the rejection region $\{T \geq C\}$ by an appropriate choice of C larger than $C_{\chi^2(q),0.05}$. Sufficient conditions when this is possible are provided in Theorems 5.1 and 6.4. Note that, in such a situation, the resulting size-controlling critical values C are then necessarily larger than $C_{\chi^2(q),0.05}$.

In light of the preceding discussion, a natural question is whether or not imposing a heteroskedasticity model more narrow than $\mathfrak{C}_{Het}$, such as

$$\mathfrak{C}_{Het,\tau_*} = \left\{\operatorname{diag}\left(\tau_1^2,\ldots,\tau_n^2\right) \in \mathfrak{C}_{Het} : \tau_i^2 \geq \tau_*^2 \text{ for all } i\right\},$$

where τ_*, $0 < \tau_* < n^{-1/2}$, is a pre-specified constant set by the user, would mitigate the failure of conventional critical values. Indeed, under the heteroskedasticity model $\mathfrak{C}_{Het,\tau_*}$, extreme concentration effects leading to rejection probabilities (arbitrarily) close to one cannot occur, and it is possible to prove that size-controlling critical values always exist when $\mathfrak{C}_{Het,\tau_*}$ is used (see Appendix A of the Supplementary Material). Unfortunately, however, this does *not* imply that conventional critical values such as $C_{\chi^2(q),0.05}$ will work. In fact, the size over $\mathfrak{C}_{Het,\tau_*}$ of tests using the critical value $C_{\chi^2(q),0.05}$ can still be considerably larger than α: to see this, observe that the sets $\mathfrak{C}_{Het,\tau_*}$ are an increasing sequence of sets as $\tau_* \downarrow 0$, the union of which is $\mathfrak{C}_{Het}$. Consequently, if τ_* is small, the size over $\mathfrak{C}_{Het,\tau_*}$ will be close to the size over $\mathfrak{C}_{Het}$, and thus the former will be much larger than α in case the latter is so. As a consequence, also in case of the more narrow heteroskedasticity model $\mathfrak{C}_{Het,\tau_*}$, size-controlling critical values larger than $C_{\chi^2(q),0.05}$ will have to be used in such a case. Furthermore, the bound τ_* has to be decided upon prior to the data analysis and is thus part of *modeling* the form of heteroskedasticity. It is difficult to see how one would come up with a reasonable value of τ_* in practice: if τ_* is chosen to be small, this may result in a

[14] This is an oversimplified description ignoring some technical details.

[15] We note, however, that there are testing problems (e.g., testing the mean in a heteroskedastic location model using the test statistic T_{uc}) for which the text-book critical values obtained under homoskedasticity are actually valid (see Bakirov and Székely, 2005). The reason is that the "worst-case" distribution in this case corresponds to homoskedasticity.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

heteroskedasticity model under which the test based on $C_{\chi^2(q),0.05}$ is still plagued by overrejection as just discussed, while choosing τ_* large will typically not be defensible as it presumes considerable knowledge about the admissible forms of heteroskedasticity.

5. SIZE-CONTROL RESULTS FOR T_{Het} AND T_{uc} WHEN $\mathfrak{C} = \mathfrak{C}_{Het}$

We introduce the following notation: for a given linear subspace $\mathcal{L}$ of $\mathbb{R}^n$, we define the set of indices $I_0(\mathcal{L})$ via

$$I_0(\mathcal{L}) = \{i : 1 \leq i \leq n, e_i(n) \in \mathcal{L}\}.$$

We set $I_1(\mathcal{L}) = \{1,\dots,n\} \backslash I_0(\mathcal{L})$. Clearly, $\operatorname{card}(I_0(\mathcal{L})) \leq \dim(\mathcal{L})$ holds. In particular, if $\dim(\mathcal{L}) < n$ holds (which, in particular, is so in the leading case $\mathcal{L} = \mathfrak{M}_0^{lin}$, since $\dim(\mathfrak{M}_0^{lin}) = k - q < n$), then $\operatorname{card}(I_0(\mathcal{L})) < n$, and thus $\operatorname{card}(I_1(\mathcal{L})) \geq 1$.

We have the following size-control result for T_{uc} as well as for T_{Het} over the heteroskedasticity model $\mathfrak{C}_{Het}$ (more precisely, over the null hypothesis H_0 described in (3) with $\mathfrak{C} = \mathfrak{C}_{Het}$). Note that $\mathfrak{C}_{Het}$ is the largest possible heteroskedasticity model and reflects complete ignorance about the form of heteroskedasticity.

THEOREM 5.1. (a) *For every* $0 < \alpha < 1$, *there exists a real number* $C(\alpha)$ *such that*

$$\sup_{\mu_0 \in \mathfrak{M}_0} \sup_{0<\sigma^2<\infty} \sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0,\sigma^2\Sigma}(T_{uc} \geq C(\alpha)) \leq \alpha \tag{7}$$

holds, provided that

$$e_i(n) \notin \operatorname{span}(X) \quad \textit{for every} \quad i \in I_1(\mathfrak{M}_0^{lin}). \tag{8}$$

Furthermore, under condition (8), even equality can be achieved in (7) by a proper choice of $C(\alpha)$, *provided* $\alpha \in (0,\alpha^*] \cap (0,1)$ *holds, where* $\alpha^* = \sup_{C\in(C^*,\infty)} \sup_{\Sigma\in\mathfrak{C}_{Het}} P_{\mu_0,\Sigma}(T_{uc} \geq C)$ *is positive and where* $C^* = \max\{T_{uc}(\mu_0 + e_i(n)) : i \in I_1(\mathfrak{M}_0^{lin})\}$ *for* $\mu_0 \in \mathfrak{M}_0$ *(with neither* α^* *nor* C^* *depending on the choice of* $\mu_0 \in \mathfrak{M}_0$*).*

(b) *Suppose Assumption 1 is satisfied.*[16] *Then, for every* $0 < \alpha < 1$, *there exists a real number* $C(\alpha)$ *such that*

$$\sup_{\mu_0 \in \mathfrak{M}_0} \sup_{0<\sigma^2<\infty} \sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0,\sigma^2\Sigma}(T_{Het} \geq C(\alpha)) \leq \alpha \tag{9}$$

holds, provided that

$$e_i(n) \notin \mathsf{B} \quad \textit{for every} \quad i \in I_1(\mathfrak{M}_0^{lin}). \tag{10}$$

[16] Condition (10) clearly implies that the set B is a proper subset of $\mathbb{R}^n$ (as $\operatorname{card}(I_1(\mathfrak{M}_0^{lin})) \geq 1$) and thus implies Assumption 1. Hence, we could have dropped this assumption from the formulation of the theorem. For clarity of presentation, we have, however, chosen to explicitly mention Assumption 1. A similar remark applies to some of the other results given below and will not be repeated.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

Furthermore, under condition (10), even equality can be achieved in (9) by a proper choice of $C(\alpha)$*, provided* $\alpha \in (0,\alpha^*] \cap (0,1)$ *holds, where now* $\alpha^* = \sup_{C\in(C^*,\infty)} \sup_{\Sigma\in\mathfrak{C}_{Het}} P_{\mu_0,\Sigma}(T_{Het} \geq C)$ *is positive and where* $C^* = \max\{T_{Het}(\mu_0 + e_i(n)) : i \in I_1(\mathfrak{M}_0^{lin})\}$ *for* $\mu_0 \in \mathfrak{M}_0$ *(with neither* α^* *nor* C^* *depending on the choice of* $\mu_0 \in \mathfrak{M}_0$*).*

(c) *Under the assumptions of Part (a) (Part (b), respectively) implying existence of a critical value* $C(\alpha)$ *satisfying (7) ((9), respectively), a smallest critical value, denoted by* $C_\diamond(\alpha)$*, satisfying (7) ((9), respectively) exists for every* $0 < \alpha < 1$*. And,* $C_\diamond(\alpha)$ *corresponding to Part (a) (Part (b), respectively) is also the smallest among the critical values leading to equality in (7) ((9), respectively) whenever such critical values exist. (Although* $C_\diamond(\alpha)$ *corresponding to Part (a) and (b), respectively, will typically be different, we use the same symbol.)*[17]

We see from the theorem that the condition for size control of T_{Het} (T_{uc}, respectively) over $\mathfrak{C}_{Het}$, i.e., condition (10) ((8), respectively), only depends on X and R; in particular, in case of T_{Het}, it does not depend on how the weights d_i figuring in the definition of T_{Het} have been chosen (note that the set B only depends on X and R). Moreover, the sufficient conditions for size control are generically satisfied in the universe of all $n \times k$ design matrices X (of rank k) (see Example 5.1 and the attending discussion further below). Furthermore, it is plain that the size-controlling critical values $C(\alpha)$ in Theorem 5.1 will depend on the choice of test statistic as well as on the testing problem at hand. More concretely, the size-controlling critical values in Part (b) of the theorem thus depend only on X, R, and r, as well as on the choice of weights d_i, whereas in Part (a), the dependence is only on X, R, and r. We do not show these dependencies in the notation. In fact, as discussed in Remark 5.2, it turns out that the size-controlling critical values in both cases actually do *not* depend on the value of r at all (provided the weights d_i are not allowed to depend on r in case of T_{Het}). Similarly, it is easy to see that C^* and α^* in Theorem 5.1 do not depend on r (under the same provision as before in case of T_{Het}).

Another observation is that any critical value delivering size control over $\mathfrak{C}_{Het}$ also delivers size control over *any* other heteroskedasticity model $\mathfrak{C}$ since $\mathfrak{C} \subseteq \mathfrak{C}_{Het}$. Of course, for such a $\mathfrak{C}$, even smaller critical values (than needed for $\mathfrak{C}_{Het}$) may already suffice for size control. Also, note that sufficient conditions implying size control over $\mathfrak{C}_{Het}$ may be more restrictive than sufficient conditions implying only size control over a smaller heteroskedasticity model $\mathfrak{C}$. For size-control results tailored to such smaller models $\mathfrak{C}$, see Appendix A of the Supplementary Material.

In light of the results of Chesher and Jewitt (1987) and Chesher (1989), it is useful to interpret the sufficient conditions for size control, i.e., (8) and (10), in terms of high-leverage points. First, note that $e_i(n) \in \operatorname{span}(X)$ is equivalent to $h_{ii} = 1$, which corresponds to the ith observation being an “extreme high-leverage point.” Hence, (8) is equivalent to $h_{ii} < 1$ for every $i \in \mathcal{I}_1(\mathfrak{M}_0^{lin})$. In other words,

[17] Cf. also Appendix A.3 of the Supplementary Material.

the condition for a size-controlling critical value to exist in Part (a) of Theorem 5.1 requires that none of the indices in $\mathcal{I}_1(\mathfrak{M}_0^{lin})$ corresponds to an extreme high-leverage point. (It is interesting to observe that all indices in $\mathcal{I}_0(\mathfrak{M}_0^{lin})$ [note that this set may be empty] correspond to extreme high-leverage points.) Hence, for the condition in (8) *not* to be satisfied, not only must extreme high-leverage points be present, but the lever needs to be of a particular type depending on the hypothesis given by (R,r) (namely, it must have $i \in \mathcal{I}_1(\mathfrak{M}_0^{lin})$). Second, note that a sufficient, but not necessary, condition for (8) is $h_{ii} < 1$, for $i = 1,\dots,n$. Sufficiency is obvious from the preceding discussion. That the condition is not necessary can be seen from Example 5.2 further below. Finally, condition (10) implies condition (8) (since $\operatorname{span}(X) \subseteq \mathsf{B}$), and hence implies $h_{ii} < 1$ for every $i \in \mathcal{I}_1(\mathfrak{M}_0^{lin})$. The converse is not always true: even $h_{ii} < 1$, for every $i = 1,\dots,n$, does not guarantee (10) to be satisfied (see Example 5.5 further below). However, *generically* (8) and (10) coincide (see Lemma A.3 in Pötscher and Preinerstorfer, 2018), in which case the discussion given above for (8) also applies to (10).

Remark 5.2 (*Independence of the value of r and implications for confidence sets*). (i) As already noted before, the sufficient conditions for size control in both parts of Theorem 5.1 only depend on X and R. In particular, they do not depend on the value of r.

(ii) The size of the test based on T_{uc} (T_{Het}, respectively) in Theorem 5.1 as well as the size-controlling critical values $C(\alpha)$ (for both test statistics) do also not depend on the value of r (provided the weights d_i are not allowed to depend on r in case of T_{Het}). This follows from Lemma 5.15 in Pötscher and Preinerstorfer (2018) combined with Remark C.1 in Appendix C of the Supplementary Material.[18] This observation is of some importance, as it allows one easily to obtain confidence sets for $R\beta$ by "inverting" the test without the need of recomputing the critical value for every value of r.

Remark 5.3 (*Some equivalencies*). If the respective smallest size-controlling critical values are used (provided they exist), the tests obtained from T_{Het} with the HC0 and the HC1 weights, respectively, are identical, as these two test statistics differ only by a multiplicative constant. The same reasoning applies to the test statistics based on the HC0–HC4 weights, respectively, in case h_{ii} does not depend on i.

Remark 5.4 (*Positivity of size-controlling critical values*). For every $0 < \alpha < 1$ any $C(\alpha)$, satisfying (7) or (9) is necessarily positive. To see this, observe that $\{T_{uc} \geq C\} = \{T_{Het} \geq C\} = \mathbb{R}^n$ for $C \leq 0$, since both test statistics are nonnegative everywhere.

The next proposition complements Theorem 5.1 and provides a useful lower bound for the size-controlling critical values (other than the trivial bound given in the preceding remark).

[18] For this argument, we impose Assumption 1 in case of T_{Het}, the case where this assumption is violated being trivial.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

Proposition 5.5. [19,20] (a) *Suppose that (8) is satisfied. Then any $C(\alpha)$ satisfying (7) necessarily has to satisfy $C(\alpha) \geq C^*$, where C^* is as in Part (a) of Theorem 5.1. In fact, for any $C < C^*$, we have $\sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0, \sigma^2 \Sigma}(T_{uc} \geq C) = 1$ for every $\mu_0 \in \mathfrak{M}_0$ and every $\sigma^2 \in (0, \infty)$.*

(b) *Suppose that Assumption 1 and (10) are satisfied. Then any $C(\alpha)$ satisfying (9) necessarily has to satisfy $C(\alpha) \geq C^*$, where C^* is as in Part (b) of Theorem 5.1. In fact, for any $C < C^*$, we have $\sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0, \sigma^2 \Sigma}(T_{Het} \geq C) = 1$ for every $\mu_0 \in \mathfrak{M}_0$ and every $\sigma^2 \in (0, \infty)$.*

The preceding observation is useful in two ways: first, critical values suggested in the literature (such as, the $(1-\alpha)$-quantile of a chi-square distribution with q degrees of freedom or critical values obtained from a degree of freedom adjustment) can immediately be dismissed if they turn out to be less than C^*, as they then certainly will not guarantee size control.[21] We use this line of reasoning in the numerical results in Section 11. Second, if the *observed* value of the test statistic T_{Het} (T_{uc}, respectively) is less than C^*, the decision not to reject the null hypothesis can be taken without further need to compute size-controlling critical values. Note that C^* as given in Theorem 5.1 is quite easy to compute in any given application.

Remark 5.6. Suppose the assumptions of Part (a) (Part (b), respectively) of Theorem 5.1 are satisfied. Then we know from that theorem that the size (over $\mathfrak{C}_{Het}$) of $\{T_{uc} \geq C_\diamond(\alpha)\}$ ($\{T_{Het} \geq C_\diamond(\alpha)\}$, respectively) equals α provided $\alpha \in (0, \alpha^*] \cap (0,1)$. If now $\alpha^* < \alpha < 1$, then the size (over $\mathfrak{C}_{Het}$) of $\{T_{uc} \geq C_\diamond(\alpha)\}$ ($\{T_{Het} \geq C_\diamond(\alpha)\}$, respectively) equals α^* (where the $C_\diamond(\alpha)$'s pertaining to Parts (a) and (b) may be different). This follows from $C_\diamond(\alpha) \geq C^*$ (see Proposition 5.5) and Remark 5.13(i) in Pötscher and Preinerstorfer (2018).[22] This argument actually also delivers that $C_\diamond(\alpha) = C^*$ must hold in case $\alpha^* < \alpha < 1$.

We next discuss to what extent the sufficient conditions for size control in Theorem 5.1 are also necessary.

[19] It is not difficult to show in the context of Parts (a) and (b) of the proposition that any critical value $C > C^*$ actually leads to size less than 1. This follows from a reasoning similar as in Remark 5.4 of Pötscher and Preinerstorfer (2018).

[20] If (10) in Part (b) of the proposition does not hold, the conclusion of Part (b) can be shown to continue to hold with C^* as defined in Theorem 5.1(b), and also with C^* as defined in Lemma 5.11 of Pötscher and Preinerstorfer (2018) (note that under the assumptions of Part (b) of the proposition both definitions of C^* actually coincide as shown in the proof of Theorem 5.1). (Recall that under violation of (10), size-controlling critical values may or may not exist.) If Assumption 1 is not satisfied, then $T_{Het} \equiv 0$, and the conclusion of Part (b) holds trivially (as $C^* = 0$ with both definitions). If (8) in Part (a) of the proposition is not satisfied, then no size-controlling critical value exists by Proposition 5.7; hence, the conclusion of Part (a) holds trivially, again regardless of which of the two definitions of C^* is adopted.

[21] In contrast, if the critical value turns out to be larger than or equal to C^*, it does *not* follow that size is less than or equal to α. In fact, substantially oversized tests using a critical value $C > C^*$ are certainly possible (see, e.g., Table 2 and the pertaining discussion).

[22] The assumptions for Part A of Proposition 5.12 in Pötscher and Preinerstorfer (2018) required in Remark 5.13 of that article are satisfied under the assumptions of Theorem 5.1 as shown in the proof of Theorem A.1 in Appendix C of the Supplementary Material. In this proof also the condition $\lambda_{\mathbb{R}^n}(T_{uc} = C^*) = 0$ ($\lambda_{\mathbb{R}^n}(T_{Het} = C^*) = 0$, respectively) required in Remark 5.13 of Pötscher and Preinerstorfer (2018) is verified.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

PROPOSITION 5.7. (a) *If (8) is violated, then* $\sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0,\sigma^2\Sigma}(T_{uc} \geq C) = 1$ *for every choice of critical value C, every* $\mu_0 \in \mathfrak{M}_0$*, and every* $\sigma^2 \in (0,\infty)$ *(implying that size equals* 1 *for every C). As a consequence, the sufficient condition for size control (8) in Part (a) of Theorem 5.1 is also necessary.*

(b) *Suppose Assumption 1 is satisfied.*[23] *If (8) is violated, then* $\sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0,\sigma^2\Sigma}$ $(T_{Het} \geq C) = 1$ *for every choice of critical value C, every* $\mu_0 \in \mathfrak{M}_0$*, and every* $\sigma^2 \in (0,\infty)$ *(implying that size equals* 1 *for every C). (In case X and R are such that* $\mathsf{B} = \operatorname{span}(X)$*, conditions (8) and (10) coincide; hence, the sufficient condition for size control (10) in Part (b) of Theorem 5.1 is then also necessary in this case.)*

Remark 5.8. Suppose Assumption 1 is satisfied. In case $\mathsf{B} \neq \operatorname{span}(X)$ and (8) hold, but (10) is violated, neither Part (b) of Theorem 5.1 nor Part (b) of Proposition 5.7 applies. We note that there are instances of this situation (see Example 5.5) for which it can be shown by other methods that T_{Het} is size-controllable despite failure of (10);[24] as a consequence, (10) is not necessary for (9) in general. We conjecture that there are other instances of the situation described here where size control is not possible, but we have not investigated this in any detail. (What can be said in general in this situation is that the size of the rejection region $\{T_{Het} \geq C\}$ over $\mathfrak{C}_{Het}$ is certainly equal to 1 for every $C < \max\{T_{Het}(\mu_0 + e_i(n)) : e_i(n) \notin \mathsf{B}\}$, where we use the convention that this maximum is $-\infty$ in case the set over which the maximum is taken is empty. This follows from Lemma 4.1 in Pötscher and Preinerstorfer (2019) with $\mathbb{K}$ equal to the collection $\{\Pi_{(\mathfrak{M}_0^{lin})^{\perp}} e_i(n) : e_i(n) \notin \mathsf{B}\}$.)

Remark 5.9. Let T stand for either T_{Het} or T_{uc}, and suppose that Assumption 1 is satisfied in case of $T = T_{Het}$: by Remark C.1 in Appendix C of the Supplementary Material and Lemma 5.16 in Pötscher and Preinerstorfer (2018), the rejection regions $\{y : T(y) \geq C\}$ and $\{y : T(y) > C\}$ differ only by a $\lambda_{\mathbb{R}^n}$-null set. Since the measures $P_{\mu,\sigma^2\Sigma}$ are absolutely continuous w.r.t. $\lambda_{\mathbb{R}^n}$ when Σ is nonsingular, $P_{\mu,\sigma^2\Sigma}(T \geq C) = P_{\mu,\sigma^2\Sigma}(T > C)$ then follows, and hence the results in this section given for rejection probabilities $P_{\mu,\sigma^2\Sigma}(T \geq C)$ apply to rejection probabilities $P_{\mu,\sigma^2\Sigma}(T > C)$ equally well (under the above provision in case of $T = T_{Het}$). A similar remark applies to the results in Appendix A.1 of the Supplementary Material.

5.1. Some Examples

We illustrate Theorem 5.1 and Proposition 5.7 with a few examples.

Example 5.1. (i) Suppose the design matrix satisfies $e_i(n) \notin \operatorname{span}(X)$ for *every* $1 \leq i \leq n$ (which will typically be the case). Then obviously the sufficient condition (8) is satisfied (in fact, for every choice of $\mathfrak{M}_0$, i.e., for every choice of restriction to be tested). And, the sufficient condition (10) is also satisfied provided $\mathsf{B} = \operatorname{span}(X)$.

[23] If this assumption is violated, then T_{Het} is identically zero, an uninteresting trivial case.

[24] In this example, actually $e_i(n) \in \mathsf{B}$ holds for all $i = 1,\ldots,n$.

(ii) Suppose the design matrix X and the restriction R are such that $e_i(n) \notin \mathsf{B}$ for *every* $1 \leq i \leq n$. Then the sufficient condition (10) is clearly satisfied.

This example shows, in particular, that the sufficient conditions for size control are generically satisfied in the universe of all $n \times k$ design matrices X (of rank k). Given the example, this is obvious for T_{uc}; and it follows for T_{Het} by additionally noting that, for every given choice of restriction to be tested, the relation $\mathsf{B} = \operatorname{span}(X)$ holds generically in the universe of all $n \times k$ design matrices X (of rank k) (see Lemma A.3 in Pötscher and Preinerstorfer, 2018). The next example discusses the case where a standard basis vector is among the regressors.

Example 5.2. Suppose that $e_1(n)$ is the first column of X and that $e_i(n) \notin \operatorname{span}(X)$ for every $2 \leq i \leq n$. Suppose further that R is of the form $R = (0, \tilde{R})$, where $\tilde{R}$ has dimension $q \times (k-1)$. That is, the restriction to be tested does not involve the coefficient of the first regressor. Then it is easy to see that (8) is satisfied and size control for T_{uc} is thus possible. If also $\mathsf{B} = \operatorname{span}(X)$ holds, then the same is true for (10) and T_{Het}. (In case R is not as above, but has a nonzero first coordinate, then it is easy to see that $1 \in I_1(\mathfrak{M}_0^{lin})$, and hence (8) is violated. It follows from Proposition 5.7 that the rejection region $\{T_{uc} \geq C\}$ indeed has size 1 for every choice of critical value C when $\mathfrak{C}_{Het}$ is the heteroskedasticity model; and the same is true for T_{Het}, provided Assumption 1 is satisfied.[25])

We continue with a few more examples where X has a particular structure.

Example 5.3. *(Heteroskedastic location model)* Suppose $k = 1$, $x_{t1} = 1$ for all t, $q = 1$, $R = 1$, and $r \in \mathbb{R}$. The heteroskedasticity model is given by $\mathfrak{C}_{Het}$. Then the conditions for size control in both parts of Theorem 5.1 are satisfied (since it is easy to see that B coincides with $\operatorname{span}(X)$ and that Assumption 1 is satisfied). Note also that in this example T_{Het} and T_{uc} actually coincide in case $d_i = n/(n-1)$ for all i, i.e., if the HC1, HC2, or HC4 weights are used, and differ only by a multiplicative constant if the HC0 or HC3 weights are employed; in particular, all these test statistics give rise to one and the same test if the respective smallest size-controlling critical values are used (cf. Remark 5.3).[26] Furthermore, note that the here observed size controllability is in line with results in Bakirov and Székely (2005) stating that, for a certain range of significance levels α, the usual critical values obtained from an $F_{1,n-1}$-distribution actually can be used as size-controlling critical values $C(\alpha)$ for the test statistic T_{uc} (in fact, these are then the smallest size-controlling critical values $C_\diamond(\alpha)$).

The subsequent example is closely related to the Behrens–Fisher problem (see Remark A.4 in Appendix A.1 of the Supplementary Material).

[25] If Assumption 1 is violated, then T_{Het} is identically zero, an uninteresting trivial case.

[26] In fact, more is true in the location model: the test statistics $\tilde{T}_{Het}$ using the HC0R–HC4R weights (defined in Section 6) all coincide (cf. Footnote 33), and they also coincide with $\tilde{T}_{uc}$ (also defined in Section 6). Perusing the connection between $\tilde{T}_{uc}$ and T_{uc} established in Section 6.2.1, we can then even conclude that all the test statistics T_{uc}, T_{Het} with HC0–HC4 weights, $\tilde{T}_{uc}$, and $\tilde{T}_{Het}$ with HC0R–HC4R weights give rise to (essentially) the same test, provided the respective smallest size-controlling critical values are used.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

Example 5.4 (*Comparing the means of two heteroskedastic groups*). Consider the problem of testing the equality of the means of two independent normal populations where the variances of each item may be different, even within a group. In our framework, this corresponds to the case $k = 2$, $x_{t1} = 1$ for $1 \leq t \leq n_1$, $x_{t1} = 0$ for $n_1 < t \leq n_1 + n_2 = n$, $x_{t2} = 1 - x_{t1}$, and $R = (1, -1)$ with $r = 0$. The heteroskedasticity model is then again $\mathfrak{C}_{Het}$. We first assume that $n_i \geq 2$ holds for $i = 1, 2$. Note that in the present context, T_{uc} is nothing else than the square of the two-sample t-statistic that uses a pooled variance estimator, and that T_{Het} is the square of the two-sample t-statistic that uses appropriate variance estimators from each group (the particular form of the variance estimator being determined by the choice of d_i). Now, $e_i(n) \notin \operatorname{span}(X)$ for *every* $1 \leq i \leq n$ holds, and hence T_{uc} is size-controllable (cf. Example 5.1(i)). This is in line with results in Bakirov (1998) (cf. also Section 5.2). Furthermore, it is obvious that Assumption 1 is satisfied (as $s = 0$) and a simple calculation shows that $B(y) = \hat{u}(y)'A$, where A is a diagonal matrix with $a_{ii} = n_1^{-1}$ for $1 \leq i \leq n_1$ and $a_{ii} = -n_2^{-1}$ else. This shows that the set B coincides with $\operatorname{span}(X)$. Consequently, also T_{Het} is size-controllable (again cf. Example 5.1(i)). We also note here that the observed size controllability of T_{Het} is in line with results in Ibragimov and Müller (2016), stating that for a certain range of significance levels α and group sizes n_i, the usual critical values obtained from an $F_{1,\,\min(n_1,n_2)-1}$ -distribution actually can be used as size-controlling critical values $C(\alpha)$ for the test statistic T_{Het} in case d_i is set equal to $(1 - h_{ii})^{-1}$; in fact, they are then the smallest size-controlling critical values (cf. the discussion preceding Theorem 1 in Ibragimov and Müller (2016)). In the rather uninteresting case $n_1 = 1$ and $n_2 \geq 2$, it is easy to see that Assumption 1 is satisfied and that the size of both tests equals 1 for all choices of critical values in view of Proposition 5.7, since $e_1(n) \in \operatorname{span}(X)$ and $1 \in I_1(\mathfrak{M}_0^{lin}) = \{1, \ldots, n\}$. The same is true if $n_1 \geq 2$ and $n_2 = 1$. (The remaining and uninteresting case $n_1 = n_2 = 1$ falls outside of our framework since we always require $n > k$.)

The next example is an extension of the previous problem to the case of more than two groups. An interesting phenomenon occurs here: the sufficient conditions for size control of T_{Het} given in Theorem 5.1 are *violated*, but size controllability can *nevertheless* be established by additional arguments. Hence, this example provides an instance where the conditions in Part (b) of Theorem 5.1 are not necessary.

Example 5.5 (*Comparing the means of k heteroskedastic groups*). We are given k integers $n_j \geq 1$ with $\sum_{j=1}^{k} n_j = n$ describing group sizes where $k \geq 3$ holds. The regressors x_{ti} for $1 \leq i \leq k$ indicate group membership, i.e., they satisfy $x_{ti} = 1$ for $\sum_{j=1}^{i-1} n_j < t \leq \sum_{j=1}^{i} n_j$ and $x_{ti} = 0$ otherwise. The heteroskedasticity model is given by $\mathfrak{C}_{Het}$. We are interested in testing $\beta_1 = \ldots = \beta_k$. We thus may choose the $(k-1) \times k$ restriction matrix R with jth row $(1, 0, \ldots, 0, -1, \ldots, 0)$ where the entry -1 is at position $j + 1$. Of course, $q = k - 1$ and $r = 0$ hold. We first consider the case where $n_j \geq 2$ for all j. Then clearly $k < n$ is satisfied. With regard to T_{uc}, we see immediately that $e_i(n) \notin \operatorname{span}(X)$ for every $1 \leq i \leq n$ follows (since $n_j \geq 2$ for all j)

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

and thus the sufficient condition (8) for size control of T_{uc} is satisfied. Turning to T_{Het}, it is easy to see that Assumption 1 is satisfied (since $s=0$ in view of $n_j \geq 2$). Furthermore, the jth row of $R(X'X)^{-1}X'$ is seen to be of the form

$$(n_1^{-1},\ldots,n_1^{-1},0,\ldots,0,-n_{j+1}^{-1},\ldots,-n_{j+1}^{-1},0\ldots,0),$$

from which it follows that

$$R(X'X)^{-1}X'\operatorname{diag}(d_1\hat{u}_1^2(y),\ldots,d_n\hat{u}_n^2(y))X(X'X)^{-1}R = S_1\iota\iota' + \operatorname{diag}(S_2,\ldots,S_k), \tag{11}$$

where ι is the $(k-1)$-dimensional vector with entries all equal to 1 and where $S_j = n_j^{-2}\sum_t d_t\hat{u}_t^2(y) = n_j^{-2}\sum_t d_t(y_t-\bar{y}_{(j)})^2$ with the summation index t running over all elements in the jth group, and where $\bar{y}_{(j)}$ is the mean in group j. From (11), it is not difficult to verify that the set B is given by

$$\mathsf{B} = \bigcup_{i,j=1,i\neq j}^{k}\left\{y\in\mathbb{R}^n : S_i(y)=S_j(y)=0\right\} = \bigcup_{i,j=1,i\neq j}^{k}\operatorname{span}\left(x_{\cdot i},x_{\cdot j},\left\{e_l(n): x_{li}=x_{lj}=0\right\}\right).$$

Note that B is not a linear space and is strictly larger than $\operatorname{span}(X)$. The set $\mathfrak{M}_0^{lin}$ is given by the span of the vector $e=(1,1,\ldots,1)'$. Hence, $I_1(\mathfrak{M}_0^{lin}) = \{1,\ldots,n\}$. Since $e_i(n)\in\mathsf{B}$ holds for every i, we conclude that the sufficient condition (10) for size control of T_{Het} is *not* satisfied and hence Part (b) of Theorem 5.1 does *not* apply. However, it can be shown by additional arguments (see Proposition C.3 in Appendix C of the Supplementary Material) that T_{Het} is nevertheless size-controllable, i.e., that (9) holds.[27] Next, in the case where $n_j=1$ for some j, but not for all j, Proposition 5.7 shows that the size of the test based on T_{uc} equals 1 for all choices of critical values, since then for some i the standard basis vector $e_i(n)$ is one of the regressors and thus we have $e_i(n)\in\operatorname{span}(X)$ and $i\in I_1(\mathfrak{M}_0^{lin})=\{1,\ldots,n\}$. For T_{Het}, the same is true if $n_j=1$ holds for exactly one j (because of Part (b) of Proposition 5.7 and since then Assumption 1 is satisfied as is easily seen); in case $n_j=1$ is true for (at least) two, but not all, values of j, T_{Het} is identically zero (as then Assumption 1 is violated), and thus is size-controllable in a trivial way. (The remaining and uninteresting case $n_j=1$ for all j falls outside of our framework since we always require $n>k$.)

We close this section by one more example. Again, the sufficient conditions in Part (b) of Theorem 5.1 fail to hold, but additional arguments based on Example 5.3 establish size controllability of the test based on T_{Het}.

Example 5.6. Consider again the situation of Example 5.4, except that now $R=I_2$, the 2×2 identity matrix (and again $r=0$). Then $q=k=2$ holds. Consider first the case where $n_i\geq 2$, for $i=1,2$. Condition (8) is then obviously satisfied, and hence T_{uc} is size-controllable. We next turn to T_{Het}. Since $\mathfrak{M}_0^{lin}=\{0\}$, we have $I_1(\mathfrak{M}_0^{lin})=\{1,\ldots,n\}$. Furthermore, simple computations show that Assumption 1

[27] A smallest size-controlling critical value then also exists in view of Appendix A.3 of the Supplementary Material.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

is satisfied and that

$$\mathsf{B} = \operatorname{span}(x_{\cdot 1}, \{e_i(n) : i > n_1\}) \cup \operatorname{span}(x_{\cdot 2}, \{e_i(n) : i \leq n_1\}).$$

Obviously, the sufficient condition (10) for size control of T_{Het} is violated. Nevertheless, T_{Het} is size-controllable by the following argument:[28] simple computations show that $T_{Het}(y) = T_1(y) + T_2(y)$ for $y \notin \mathsf{B}$, where $T_1(y) = n_1^2 \hat{\beta}_1^2(y) / \sum_{t=1}^{n_1} d_t \hat{u}_t^2(y)$ and $T_2(y) = n_2^2 \hat{\beta}_2^2(y) / \sum_{t=n_1+1}^{n} d_t \hat{u}_t^2(y)$. (If the denominator in the formula for $T_i(y)$ is zero for some $y \in \mathbb{R}^n$, we define $T_i(y)$ as zero.) Since B is a $\lambda_{\mathbb{R}^n}$-null set, $P_{0,\sigma^2\Sigma}(T_{Het} \geq C) \leq P_{0,\sigma^2\Sigma}(T_1 \geq C/2) + P_{0,\sigma^2\Sigma}(T_2 \geq C/2)$ for $C > 0$. Now, it is easy to see that $P_{0,\sigma^2\Sigma}(T_i \geq C/2)$, for $i = 1, 2$, coincides with the null rejection probability of a test for the mean in a heteroskedastic location model (based on a test statistic of the form (4)). However, as shown in Example 5.3, such a test is size-controllable. (In the case $n_1 = 1$ and $n_2 \geq 2$ (or vice versa), condition (8) is violated and the rejection region $\{T_{uc} \geq C\}$ has size 1 for every C; furthermore, Assumption 1 is violated, and hence T_{Het} is identically zero. The case $n_1 = n_2 = 1$ falls outside of our framework as then $k = n$.)

In Appendix C of the Supplementary Material, we discuss yet another example where the sufficient condition of Part (b) of Theorem 5.1 fails, but size controllability can nevertheless be established.

5.2. Some Variations on Bakirov and Székely (2005)

(i) As noted in Ibragimov and Müller (2010), testing a hypothesis regarding a *scalar* linear contrast in a heteroskedastic (Gaussian) linear regression model more general than a location model can often be converted to a testing problem in a heteroskedastic (Gaussian) location model by suitably dividing the data into subgroups and by considering groupwise least-squares estimators, thus making it amenable to the Bakirov and Székely (2005) result mentioned in Section 1. However, this introduces additional questions such as how to divide up the data. In any case, this approach is limited to testing hypotheses on *scalar* linear contrasts. It also requires that the linear contrast subject to test is *estimable* in each subgroup.

(ii) In case the linear contrast subject to test is *not* estimable in each subgroup, but can be written as the difference of two linear contrasts where the first contrast is estimable in the first G_1 groups whereas the second contrast is estimable in the last G_2 groups (where we consider a total of $G_1 + G_2$ groups), Ibragimov and Müller (2016) point out that the problem can be converted into the problem of comparing two heteroskedastic (Gaussian) populations. Now, for such a two-sample comparison problem, Bakirov (1998) shows for a certain two-sample t-statistic (the square of which is T_{uc}; cf. Example 5.4) how—in the presence of heteroskedasticity—size-controlling critical values can be constructed by appropriately transforming quantiles of a t-distribution; this result imposes conditions

[28] A smallest size-controlling critical value then also exists in view of Appendix A.3 of the Supplementary Material.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

which entail that the nominal significance level α must be quite small (requiring α not to exceed 0.01 for many group sizes, and often to be considerably smaller). This somewhat limits the applicability of Bakirov's result. Thus, Ibragimov and Müller (2016) go on to consider another two-sample t-statistic (the square of which is T_{Het} with $d_i = (1 - h_{ii})^{-1}$; cf. Example 5.4) and—extending a result in Mickey and Brown (1966)—provide a Bakirov and Székely (2005)-type result, i.e., they show that the $(1 - \alpha/2)$-quantile of a t-distribution with degrees of freedom equal to the smaller of the two sample sizes minus 1 provides the smallest size-controlling critical value (for the two-sided test) even under heteroskedasticity.[29] This result holds under certain conditions on the sample sizes and only for small α, but, e.g., allows for the choice $\alpha = 0.05$. (We note here that the description of Bakirov's (1998) result in Ibragimov and Müller (2016) is inaccurate in that a certain transformation of the critical value is being ignored.)

(iii) In the problem of comparing two heteroskedastic (Gaussian) populations based on samples of equal size ("balanced design") one can—instead of using the two-sample t -test statistics considered in Bakirov (1998) and Ibragimov and Müller (2016)—employ the Bartlett test statistic, which simply is the usual t-test statistic computed from the differences between the observations in the two samples.[30] An advantage of this approach is that the original Bakirov and Székely (2005) result is directly applicable, and there is no need to resort to the results described in (ii).

(iv) Another quite special case that can be brought under the realm of the Bakirov and Székely (2005) result is a heteroskedastic (Gaussian) regression model with only one regressor that never takes the value zero. Dividing the tth equation in the regression model by x_t converts this into a heteroskedastic location problem.

(v) The results in (i)–(iv) immediately also apply if the errors in the regression are distributed as scale mixtures of Gaussians (cf. also Section 7.1).

6. RESULTS FOR HETEROSKEDASTICITY ROBUST TEST STATISTICS USING RESTRICTED RESIDUALS

In this section, we consider two further test statistics which are versions of T_{Het} and T_{uc} with the only difference that the covariance matrix estimators used are based on restricted—instead of unrestricted—residuals. The first one of these test statistics has been suggested in the literature, e.g., in Davidson and MacKinnon (1985). We thus define

$$\tilde{T}_{Het}(y) = \begin{cases} (R\hat{\beta}(y) - r)'\tilde{\Omega}_{Het}^{-1}(y)(R\hat{\beta}(y) - r), & \text{if } \det \tilde{\Omega}_{Het}(y) \neq 0, \\ 0, & \text{if } \det \tilde{\Omega}_{Het}(y) = 0, \end{cases} \tag{12}$$

[29] In the balanced case (i.e., if the two samples have the same cardinality), the test statistic considered in Bakirov (1998) actually coincides with the test statistic in Ibragimov and Müller (2016).

[30] Certainly, there is some arbitrariness in how the observations are being "paired."

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

where $\tilde{\Omega}_{Het} = R\tilde{\Psi}_{Het}R'$ and where $\tilde{\Psi}_{Het}$ is given by

$$\tilde{\Psi}_{Het}(y) = (X'X)^{-1}X'\operatorname{diag}\left(\tilde{d}_1\tilde{u}_1^2(y), \ldots, \tilde{d}_n\tilde{u}_n^2(y)\right)X(X'X)^{-1},$$

where the constants $\tilde{d}_i > 0$ sometimes depend on the design matrix and on the restriction matrix R. Here, $\tilde{u}(y) = y - X\tilde{\beta}_{\mathfrak{M}_0}(y) = \Pi_{(\mathfrak{M}_0^{lin})^\perp}(y - \mu_0)$, where the last expression does not depend on the choice of $\mu_0 \in \mathfrak{M}_0$, and where $\tilde{u}_t(y)$ denotes the tth component of $\tilde{u}(y)$. Typical choices for $\tilde{d}_i$ are $\tilde{d}_i = 1$, $\tilde{d}_i = n/(n-(k-q))$, $\tilde{d}_i = (1-\tilde{h}_{ii})^{-1}$, or $\tilde{d}_i = (1-\tilde{h}_{ii})^{-2}$, where $\tilde{h}_{ii}$ denotes the ith diagonal element of the projection matrix $\Pi_{\mathfrak{M}_0^{lin}}$ (see, e.g., Davidson and MacKinnon, 1985). Another suggestion is $\tilde{d}_i = (1-\tilde{h}_{ii})^{-\tilde{\delta}_i}$ for $\tilde{\delta}_i = \min(n\tilde{h}_{ii}/(k-q), 4)$ with the convention that $\tilde{\delta}_i = 0$ if $k = q$.[31] For the last three choices of $\tilde{d}_i$ just given, we use the convention that we set $\tilde{d}_i = 1$ in case $\tilde{h}_{ii} = 1$. Note that $\tilde{h}_{ii} = 1$ implies $\tilde{u}_i(y) = 0$ for every y, and hence it is irrelevant which real value is assigned to $\tilde{d}_i$ in case $\tilde{h}_{ii} = 1$.[32] The five examples for the weights $\tilde{d}_i$ just given correspond to what is often called HC0R–HC4R weights in the literature.[33]

The subsequent assumption ensures that the set of y's for which $\tilde{\Omega}_{Het}(y)$ is singular is a Lebesgue null set, implying that our choice of assigning $\tilde{T}_{Het}(y)$ the value zero in case $\tilde{\Omega}_{Het}(y)$ is singular has no import on the probabilistic results of the article (as the measures $P_{\mu,\sigma^2\Sigma}$ are absolutely continuous). Also, as discussed further below, the assumption is in a certain sense unavoidable when using $\tilde{T}_{Het}$.

Assumption 2. Let $1 \leq i_1 < \cdots < i_s \leq n$ denote all the indices for which $e_{i_j}(n) \in \mathfrak{M}_0^{lin}$ holds where $e_j(n)$ denotes the jth standard basis vector in $\mathbb{R}^n$. If no such index exists, set $s = 0$. Let $X'(\neg(i_1, \ldots i_s))$ denote the matrix which is obtained from X' by deleting all columns with indices i_j, $1 \leq i_1 < \cdots < i_s \leq n$ (if $s = 0$, no column is deleted). Then $\operatorname{rank}\left(R(X'X)^{-1}X'(\neg(i_1, \ldots, i_s))\right) = q$ holds.

Observe that this assumption only depends on X and R and hence can be checked. Obviously, a simple sufficient condition for Assumption 2 to hold is that $s = 0$ (i.e., that $e_j(n) \notin \mathfrak{M}_0^{lin}$ for all j), a generically satisfied condition. Furthermore, we introduce the matrix

$$\begin{aligned}\tilde{B}(y) &= R(X'X)^{-1}X'\operatorname{diag}(\tilde{u}_1(y), \ldots, \tilde{u}_n(y)) \\ &= R(X'X)^{-1}X'\operatorname{diag}\left(e_1'(n)\Pi_{(\mathfrak{M}_0^{lin})^\perp}(y-\mu_0), \ldots, e_n'(n)\Pi_{(\mathfrak{M}_0^{lin})^\perp}(y-\mu_0)\right).\end{aligned} \tag{13}$$

Note that this matrix does not depend on the choice of $\mu_0 \in \mathfrak{M}_0$. The following lemma collects some important properties of $\tilde{\Omega}_{Het}$ and $\tilde{\mathsf{B}}$ (defined in that lemma) and is reproduced from Pötscher and Preinerstorfer (2023) for ease of reference.

[31] Note that in case $k = q$, we have $\tilde{h}_{ii} = 0$, and hence $\tilde{d}_i = 1$ regardless of our convention for $\tilde{\delta}_i$.

[32] In fact, $\tilde{h}_{ii} = 1$ is equivalent to $\tilde{u}_i(y) = 0$ for every y, each of which in turn is equivalent to $e_i(n) \in \mathfrak{M}_0^{lin}$.

[33] In the case $k = q$, the HC0R–HC4R weights all coincide ($\tilde{d}_i = 1$ for every i), and hence result in the same test statistic.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

Lemma 6.1.

(a) $\tilde{\Omega}_{Het}(y)$ *is nonnegative definite for every* $y \in \mathbb{R}^n$.
(b) $\tilde{\Omega}_{Het}(y)$ *is singular (zero, respectively) if and only if* $\operatorname{rank}(\tilde{B}(y)) < q$ $(\tilde{B}(y) = 0$, *respectively).*
(c) *The set* $\tilde{\mathsf{B}}$ *given by* $\{y \in \mathbb{R}^n : \operatorname{rank}(\tilde{B}(y)) < q\}$ *(or, in view of (b), equivalently given by* $\{y \in \mathbb{R}^n : \det(\tilde{\Omega}_{Het}(y)) = 0\}$*) is either a* $\lambda_{\mathbb{R}^n}$*-null set or the entire sample space* $\mathbb{R}^n$*. The latter occurs if and only if Assumption 2 is violated (in which case, the test based on* $\tilde{T}_{Het}$ *becomes trivial, as then* $\tilde{T}_{Het}$ *is identically zero).*
(d) *Suppose Assumption 2 holds. Then, for every* $\mu_0 \in \mathfrak{M}_0$*, the set* $\tilde{\mathsf{B}} - \mu_0$ *is a finite union of proper linear subspaces; in case* $q = 1$*,* $\tilde{\mathsf{B}} - \mu_0$ *is even a proper linear subspace itself.*[34,35] *(Note that* $\tilde{\mathsf{B}} - \mu_0$ *does not depend on the choice of* $\mu_0 \in \mathfrak{M}_0$*. In particular, if* $r = 0$*, i.e., if* $\mathfrak{M}_0$ *is linear, we thus may set* $\mu_0 = 0$*.)*
(e) $\tilde{\mathsf{B}}$ *is a closed set and contains* $\mathfrak{M}_0$*. Also,* $\tilde{\mathsf{B}}$ *is* $G(\mathfrak{M}_0)$*-invariant, and in particular,* $\tilde{\mathsf{B}} + \mathfrak{M}_0^{lin} = \tilde{\mathsf{B}}$.

In light of Part (c) of the lemma, we see that Assumption 2 is a natural and unavoidable condition if one wants to obtain a sensible test from $\tilde{T}_{Het}$.[36] Furthermore, note that if $\tilde{\mathsf{B}} = \mathfrak{M}_0$ is true, then Assumption 2 must be satisfied (since $\mathfrak{M}_0$ is a $\lambda_{\mathbb{R}^n}$-null set as $k - q < n$ is always the case). For later use, we also mention that under Assumption 2, the statistic $\tilde{T}_{Het}$ is continuous at every $y \in \mathbb{R}^n \backslash \tilde{\mathsf{B}}$.[37]

We finally consider for completeness, and in analogy with T_{uc},

$$\tilde{T}_{uc}(y) = \begin{cases} (R\hat{\beta}(y) - r)' \left(\tilde{\sigma}^2(y) R \left(X'X\right)^{-1} R'\right)^{-1} (R\hat{\beta}(y) - r), & \text{if } y \notin \mathfrak{M}_0, \\ 0, & \text{if } y \in \mathfrak{M}_0, \end{cases} \tag{14}$$

where $\tilde{\sigma}^2(y) = \tilde{u}(y)'\tilde{u}(y)/(n-(k-q)) \geq 0$ (which vanishes if and only if $y \in \mathfrak{M}_0$). Of course, our choice to set $\tilde{T}_{uc}(y) = 0$ for $y \in \mathfrak{M}_0$ again has no import on the probabilistic results in the article, since $\mathfrak{M}_0$ is a $\lambda_{\mathbb{R}^n}$-null set (and since the measures $P_{\mu,\sigma^2\Sigma}$ are absolutely continuous). For later use, we also mention that $\tilde{T}_{uc}$ is continuous at every $y \in \mathbb{R}^n \backslash \mathfrak{M}_0$. As we shall see in Section 6.2.1, there is a close connection between $\tilde{T}_{uc}$ and T_{uc}.

Remark 6.2. The test statistics $\tilde{T}_{Het}$ as well as $\tilde{T}_{uc}$ are $G(\mathfrak{M}_0)$-invariant as is easily seen (with the respective exceptional sets $\tilde{\mathsf{B}}$ and $\mathfrak{M}_0$ also being $G(\mathfrak{M}_0)$-invariant), but typically they are *not* nonsphericity-corrected F-type tests in the sense of Section 5.4 in Preinerstorfer and Pötscher (2016).

[34] Consequently, $\tilde{\mathsf{B}}$ is a finite union of proper affine subspaces, and is a proper affine subspace itself in case $q = 1$.

[35] If Assumption 2 is violated, then $\tilde{\mathsf{B}} - \mu_0 = \tilde{\mathsf{B}} = \mathbb{R}^n$ in view of Part (c).

[36] If this assumption is violated, then $\tilde{T}_{Het}$ is identically zero, an uninteresting trivial case.

[37] If Assumption 2 is violated, then $\tilde{T}_{Het}$ is constant equal to zero, and hence trivially continuous everywhere.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

Remark 6.3. Remark 3.3 also applies to $\tilde{T}_{Het}$ and $\tilde{T}_{uc}$. (To see this, note that the respective exceptional sets $\tilde{\mathsf{B}}$ and $\mathfrak{M}_0$ are the same irrespective of whether (R,r) or $(\bar{R},\bar{r})$ is used, and that A cancels out in the respective quadratic forms appearing in the definitions of the test statistics.)

6.1. Size-Control Results for $\tilde{T}_{Het}$ and $\tilde{T}_{uc}$ when $\mathfrak{C} = \mathfrak{C}_{Het}$

Here, we discuss size-control results for $\tilde{T}_{uc}$ as well as for $\tilde{T}_{Het}$ over the heteroskedasticity model $\mathfrak{C}_{Het}$ (more precisely, over the null hypothesis H_0 described in (3) with $\mathfrak{C} = \mathfrak{C}_{Het}$). Some peculiar properties of the test statistics $\tilde{T}_{uc}$ and $\tilde{T}_{Het}$ are then discussed in the following section.

We note that the first statement in Part (a) of the subsequent theorem is actually trivial, since $\tilde{T}_{uc}$ is bounded as shown in the next section (which also provides a discussion when nontrivial size-controlling critical values exist).

THEOREM 6.4. (a) *For every* $0 < \alpha < 1$, *there exists a real number* $C(\alpha)$ *such that*

$$\sup_{\mu_0 \in \mathfrak{M}_0} \sup_{0<\sigma^2<\infty} \sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0,\sigma^2\Sigma}(\tilde{T}_{uc} \geq C(\alpha)) \leq \alpha \tag{15}$$

holds. Furthermore, even equality can be achieved in (15) by a proper choice of $C(\alpha)$, *provided* $\alpha \in (0,\alpha^*] \cap (0,1)$ *holds, where* $\alpha^* = \sup_{C \in (C^*,\infty)} \sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0,\Sigma}$ $(\tilde{T}_{uc} \geq C)$ *and where* $C^* = \max\{\tilde{T}_{uc}(\mu_0 + e_i(n)) : i \in I_1(\mathfrak{M}_0^{lin})\}$ *for* $\mu_0 \in \mathfrak{M}_0$ *(with neither* α^* *nor* C^* *depending on the choice of* $\mu_0 \in \mathfrak{M}_0$*).*

(b) *Suppose Assumption 2 is satisfied.*[38] *Suppose further that* $\tilde{T}_{Het}$ *is not constant on* $\mathbb{R}^n \backslash \tilde{\mathsf{B}}$.[39] *Then, for every* $0 < \alpha < 1$, *there exists a real number* $C(\alpha)$ *such that*

$$\sup_{\mu_0 \in \mathfrak{M}_0} \sup_{0<\sigma^2<\infty} \sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0,\sigma^2\Sigma}(\tilde{T}_{Het} \geq C(\alpha)) \leq \alpha \tag{16}$$

holds, provided that for some $\mu_0 \in \mathfrak{M}_0$ *(and hence for all* $\mu_0 \in \mathfrak{M}_0$*),*

$$\mu_0 + e_i(n) \notin \tilde{\mathsf{B}} \quad \text{for every} \quad i \in I_1(\mathfrak{M}_0^{lin}). \tag{17}$$

Furthermore, under condition (17), even equality can be achieved in (16) by a proper choice of $C(\alpha)$, *provided* $\alpha \in (0,\alpha^*] \cap (0,1)$ *holds, where now* $\alpha^* = \sup_{C \in (C^*,\infty)} \sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0,\Sigma}(\tilde{T}_{Het} \geq C)$ *and where* $C^* = \max\{\tilde{T}_{Het}(\mu_0 + e_i(n)) :$

[38] Condition (17) clearly implies that the set $\tilde{\mathsf{B}}$ is a proper subset of $\mathbb{R}^n$ and thus implies Assumption 2. Hence, we could have dropped this assumption from the formulation of the theorem. A similar remark applies to some of the other results given below and will not be repeated.

[39] The case where $\tilde{T}_{Het}$ is constant on $\mathbb{R}^n \backslash \tilde{\mathsf{B}}$ can actually occur under Assumption 2 (see Remark D.2 in Appendix D of the Supplementary Material). In such a case, $\tilde{T}_{Het}$ is trivially size-controllable (since $\tilde{\mathsf{B}}$ is a $\lambda_{\mathbb{R}^n}$-null set under Assumption 2 and since all probability measures in (2) are absolutely continuous). However, neither a smallest size-controlling critical value exists (when considering rejection regions of the form $\{\tilde{T}_{Het} \geq C\}$) nor can exact size controllability be achieved for $0 < \alpha < 1$. (If Assumption 2 is violated, $\tilde{T}_{Het}$ is identically zero and a similar remark applies.)

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

$i \in I_1(\mathfrak{M}_0^{lin})\}$ for $\mu_0 \in \mathfrak{M}_0$ (with neither α^* nor C^* depending on the choice of $\mu_0 \in \mathfrak{M}_0$).

(c) *Under the assumptions of Part (a) (Part (b), respectively) implying existence of a critical value $C(\alpha)$ satisfying (15) ((16), respectively), a smallest critical value, denoted by $C_\diamond(\alpha)$, satisfying (15) ((16), respectively) exists for every* $0 < \alpha < 1$.[40] *And, $C_\diamond(\alpha)$ corresponding to Part (a) (Part (b), respectively) is also the smallest among the critical values leading to equality in (15) ((16), respectively) whenever such critical values exist. (Although $C_\diamond(\alpha)$ corresponding to Part (a) and (b), respectively, will typically be different, we use the same symbol.)*[41]

We see from the theorem that $\tilde{T}_{uc}$ is always size-controllable over $\mathfrak{C}_{Het}$, but as discussed in Section 6.2, there is a caveat: Unless (8), i.e., the necessary and sufficient condition for size controllability of T_{uc}, is satisfied, size-controlling $\tilde{T}_{uc}$ leads to trivial tests. We also see that the condition for size control of $\tilde{T}_{Het}$ over $\mathfrak{C}_{Het}$, i.e., condition (17) is always satisfied in case $\tilde{\mathsf{B}} = \mathfrak{M}_0$ (since (17) is then equivalent to $e_i(n) \notin \mathfrak{M}_0^{lin}$ for every $i \in I_1(\mathfrak{M}_0^{lin})$). Furthermore, condition (17) always only depends on X and R; in particular, it does not depend on how the weights $\tilde{d}_i$ figuring in the definition of $\tilde{T}_{Het}$ have been chosen (note that $\mu_0 + e_i(n) \notin \tilde{\mathsf{B}}$ is equivalent to $e_i(n) \notin \tilde{\mathsf{B}} - \mu_0$ and that the set $\tilde{\mathsf{B}} - \mu_0$ depends only on X and R). Furthermore, the size-controlling critical values $C(\alpha)$ in Part (b) of the preceding theorem depend only on X, R, and r, as well as on the choice of weights $\tilde{d}_i$, whereas in Part (a) the dependence is only on X, R, and r. We do not show these dependencies in the notation. In fact, as shown in Lemma D.3 in Appendix D of the Supplementary Material, it turns out that the size and the size-controlling critical values in both cases actually do *not* depend on the value of r at all (provided the weights $\tilde{d}_i$ are not allowed to depend on r in case of $\tilde{T}_{Het}$). Similarly, it is easy to see that α^* and C^* do not depend on r (under the same provision as before in case of $\tilde{T}_{Het}$).

Similarly as in Section 5, a critical value delivering size control over $\mathfrak{C}_{Het}$ also delivers size control over *any* other heteroskedasticity model $\mathfrak{C}$ since $\mathfrak{C} \subseteq \mathfrak{C}_{Het}$. Of course, for such a $\mathfrak{C}$, even smaller critical values (than needed for $\mathfrak{C}_{Het}$) may already suffice for size control. Also, note that sufficient conditions implying size control over $\mathfrak{C}_{Het}$ may be more restrictive than sufficient conditions only implying size control over a smaller heteroskedasticity model $\mathfrak{C}$. For size-control results tailored to such smaller models $\mathfrak{C}$, see Appendix A of the Supplementary Material.

Remark 6.5 (*Some equivalencies*). If the respective smallest size-controlling critical values are used (provided they exist), the tests obtained from $\tilde{T}_{Het}$ with the HC0R and the HC1R weights, respectively, are identical, as these two test statistics differ only by a multiplicative constant. The same reasoning applies to the test statistics based on the HC0R–HC4R weights, respectively, in case $\tilde{h}_{ii}$ does not depend on i.

[40] Note that there are in fact no assumptions for Part (a). We have chosen this formulation for reasons of brevity.

[41] Cf. also Appendix A.3 of the Supplementary Material.

Remark 6.6 (*Positivity of size-controlling critical values*). For every $0<\alpha<1$, any $C(\alpha)$ satisfying (15) or (16) is necessarily positive. To see this, observe that $\{\tilde{T}_{uc}\geq C\}=\{\tilde{T}_{Het}\geq C\}=\mathbb{R}^n$ for $C\leq 0$, since both test statistics are nonnegative everywhere.

The next proposition complements Theorem 6.4 and provides a lower bound for the size-controlling critical values (other than the trivial bound given in the preceding remark). The lower bound is useful for the same reasons as discussed subsequent to Proposition 5.5.

PROPOSITION 6.7. [42,43] (a) *Any $C(\alpha)$ satisfying (15) necessarily has to satisfy $C(\alpha)\geq C^*$, where C^* is as in Part (a) of Theorem 6.4. In fact, for any $C<C^*$, we have $\sup_{\Sigma\in\mathfrak{C}_{Het}}P_{\mu_0,\sigma^2\Sigma}(\tilde{T}_{uc}\geq C)=1$ for every $\mu_0\in\mathfrak{M}_0$ and every $\sigma^2\in(0,\infty)$.*

(b) *Suppose Assumption 2 and (17) are satisfied, and that $\tilde{T}_{Het}$ is not constant on $\mathbb{R}^n\backslash\tilde{\mathsf{B}}$. Then any $C(\alpha)$ satisfying (16) necessarily has to satisfy $C(\alpha)\geq C^*$, where C^* is as in Part (b) of Theorem 6.4. In fact, for any $C<C^*$, we have $\sup_{\Sigma\in\mathfrak{C}_{Het}}P_{\mu_0,\sigma^2\Sigma}(\tilde{T}_{Het}\geq C)=1$ for every $\mu_0\in\mathfrak{M}_0$ and every $\sigma^2\in(0,\infty)$.*

Remark 6.8. Suppose the assumptions of Part (a) (Part (b), respectively) of Theorem 6.4 are satisfied. Then we know from that theorem that the size (over $\mathfrak{C}_{Het}$) of $\{\tilde{T}_{uc}\geq C_\diamond(\alpha)\}$ ($\{\tilde{T}_{Het}\geq C_\diamond(\alpha)\}$, respectively) equals α provided $\alpha\in(0,\alpha^*]\cap(0,1)$. If now $\alpha^*<\alpha<1$, then the size (over $\mathfrak{C}_{Het}$) of $\{\tilde{T}_{uc}\geq C_\diamond(\alpha)\}$ ($\{\tilde{T}_{Het}\geq C_\diamond(\alpha)\}$, respectively) equals α^* (where the $C_\diamond(\alpha)$'s pertaining to Parts (a) and (b) may be different). This follows from $C_\diamond(\alpha)\geq C^*$ (see Proposition 6.7) and Remark 5.13(i) in Pötscher and Preinerstorfer (2018)).[44] This argument actually also delivers that $C_\diamond(\alpha)=C^*$ must hold in case $\alpha^*<\alpha<1$.

Remark 6.9. In contrast to Section 5, we have little information on the extent to which the sufficient conditions for size control in Part (b) of Theorem 6.4 are also necessary. This is due to the fact that $\tilde{T}_{Het}$ is typically not a nonsphericity-corrected F-type test as noted in Remark 6.2. What can be said in general in the context of Part (b) of Theorem 6.4 in case (17) is violated, is that the size of the rejection region $\{\tilde{T}_{Het}\geq C\}$ over $\mathfrak{C}_{Het}$ is certainly equal to 1 for every $C<\max\{\tilde{T}_{Het}(\mu_0+e_i(n)):\mu_0+e_i(n)\notin\tilde{\mathsf{B}}\}$, where $\mu_0\in\mathfrak{M}_0$ is arbitrary (the maximum being independent of the choice of $\mu_0\in\mathfrak{M}_0$) and where we use the

[42] It is not difficult to show in the context of Parts (a) and (b) of the proposition that any critical value $C>C^*$ actually leads to size less than 1. This follows from a reasoning similar as in Remark 5.4 of Pötscher and Preinerstorfer (2018).

[43] If (17) in Part (b) of the proposition does not hold, the conclusion of Part (b) can be shown to continue to hold with C^* as defined in Theorem 6.4(b), and also with C^* as defined in Lemma 5.11 of Pötscher and Preinerstorfer (2018) (note that under the assumptions of Part (b) of the proposition, both definitions of C^* actually coincide as shown in the proof of Theorem 6.4). If $\tilde{T}_{Het}$ is constant on $\mathbb{R}^n\backslash\tilde{\mathsf{B}}$ or if Assumption 2 fails (the latter implying $\tilde{T}_{Het}\equiv 0$), the conclusion of Part (b) also holds as is easily seen (regardless of which of the two definitions of C^* is adopted).

[44] The assumptions for Part A of Proposition 5.12 in Pötscher and Preinerstorfer (2018) required in Remark 5.13 of that article are satisfied under the assumptions of Theorem 6.4 as shown in the proof of Theorem A.5 in Appendix D of the Supplementary Material. In this proof also the condition $\lambda_{\mathbb{R}^n}(\tilde{T}_{uc}=C^*)=0$ ($\lambda_{\mathbb{R}^n}(\tilde{T}_{Het}=C^*)=0$, respectively) required in Remark 5.13 of Pötscher and Preinerstorfer (2018) is verified.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

convention that this maximum is $-\infty$ in case the set over which the maximum is taken is empty. This follows from Lemma 4.1 in Pötscher and Preinerstorfer (2019) with $\mathbb{K}$ equal to the collection $\{\Pi_{(\mathfrak{M}_0^{lin})^\perp} e_i(n) : \mu_0 + e_i(n) \notin \tilde{\mathsf{B}}\}$.

Remark 6.10. Suppose $q = k$. Then Assumption 2 is always satisfied (since $\mathfrak{M}_0$ being a singleton $\{\mu_0\}$ implies $\mathfrak{M}_0^{lin} = \{0\}$, and thus $s = 0$ in Assumption 2). The subsequent claims are proved in Appendix D of the Supplementary Material.

(i) In case $q = k > 1$, it is not difficult to see that then $\mu_0 + e_i(n) \in \tilde{\mathsf{B}}$, for every $i = 1, \ldots, n$, holds, implying that the sufficient condition (17) in Theorem 6.4(b) is violated. (In contrast, in case $q = k = 1$, both examples where (17) is satisfied as well as examples where (17) is not satisfied can be found.)

(ii) Despite of (i), in case $q = k \geq 1$, the test statistic $\tilde{T}_{Het}$ is always size-controllable over $\mathfrak{C}_{Het}$. This is so since in case $q = k \geq 1$ the statistic $\tilde{T}_{Het}$ is a bounded function.

(iii) We also note that in case $q = k \geq 1$, both the case where $\tilde{T}_{Het}$ is constant on $\mathbb{R}^n \backslash \tilde{\mathsf{B}}$ as well as the case where $\tilde{T}_{Het}$ is not constant on $\mathbb{R}^n \backslash \tilde{\mathsf{B}}$ can occur. (In the latter case, a *smallest* size-controlling critical value exists in view of Appendix A.3 of the Supplementary Material. In the former case, no *smallest* size-controlling critical value exists [when considering rejection regions of the form $\{\tilde{T}_{Het} \geq C\}$].)

Remark 6.11. Let $\tilde{T}$ stand for either $\tilde{T}_{Het}$ or $\tilde{T}_{uc}$, where in case of $\tilde{T} = \tilde{T}_{Het}$ we suppose that Assumption 2 is satisfied and that $\tilde{T}_{Het}$ is not constant on $\mathbb{R}^n \backslash \tilde{\mathsf{B}}$: By Lemma D.1 in Appendix D of the Supplementary Material, the rejection regions $\{y : \tilde{T}(y) \geq C\}$ and $\{y : \tilde{T}(y) > C\}$ differ only by a $\lambda_{\mathbb{R}^n}$-null set. Since the measures $P_{\mu,\sigma^2\Sigma}$ are absolutely continuous w.r.t. $\lambda_{\mathbb{R}^n}$ when Σ is nonsingular, $P_{\mu,\sigma^2\Sigma}(\tilde{T} \geq C) = P_{\mu,\sigma^2\Sigma}(\tilde{T} > C)$ then follows, and hence the results in this and the subsequent section given for rejection probabilities $P_{\mu,\sigma^2\Sigma}(\tilde{T} \geq C)$ apply to rejection probabilities $P_{\mu,\sigma^2\Sigma}(\tilde{T} > C)$ equally well (under the above provision in case of $T = \tilde{T}_{Het}$). A similar remark applies to the results in Appendix A.2 of the Supplementary Material.

6.2. Tests Obtained from $\tilde{T}_{uc}$ or $\tilde{T}_{Het}$ Can Be Trivial

For the test statistic T_{uc}, the rejection regions $\{T_{uc} \geq C\}$, as well as their complements, have positive (n-dimensional) Lebesgue measure for every positive real number C.[45] This follows from Parts 5 and 6 of Lemma 5.15 in Preinerstorfer and Pötscher (2016) together with Remark C.1 in Appendix C of the Supplementary Material. As a consequence, all rejection probabilities—under the null as well as under the alternative—are positive and less than one regardless of the choice of $C > 0$. (This is so because of our Gaussianity assumption and the fact that

[45] The case $C \leq 0$ is uninteresting as the rejection region of T_{uc} (and of all other test statistics considered) then are the entire space $\mathbb{R}^n$, since T_{uc} (and the other test statistics considered) take on only nonnegative values.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

all $\Sigma \in \mathfrak{C}_{Het}$ are positive definite.) For similar reasons, the same is true for T_{Het} provided Assumption 1 is satisfied.[46] The situation is somewhat different for tests derived from $\tilde{T}_{uc}$ or $\tilde{T}_{Het}$ as we shall discuss next. In the course of this, we also establish a connection between T_{uc} and $\tilde{T}_{uc}$ that is of independent interest. In this section, the size of a test always refers to size over $\mathfrak{C}_{Het}$.

6.2.1. *The Case of* $\tilde{T}_{uc}$. First, observe that $\tilde{T}_{uc}(y) \leq n-(k-q)$ holds for every $y \in \mathbb{R}^n$ and that this bound is sharp. To see this, note that using standard least-squares theory,

$$\tilde{T}_{uc}(y) = (n-(k-q))\left(1 - \sum_{i=1}^{n} \hat{u}_i^2(y) / \sum_{i=1}^{n} \tilde{u}_i^2(y)\right) \leq n-(k-q) \qquad \textbf{(18)}$$

for $y \notin \mathfrak{M}_0$ and that $\tilde{T}_{uc}(y) = 0$ else; the bound is attained precisely for $y \in \operatorname{span}(X) \backslash \mathfrak{M}_0$. An immediate consequence of this observation is that any critical value $C \geq (n-(k-q))$ leads to a test with rejection region $\{\tilde{T}_{uc} \geq C\}$ that is either empty (if $C > n-(k-q)$) or is a $\lambda_{\mathbb{R}^n}$-null set, namely $\operatorname{span}(X) \backslash \mathfrak{M}_0$ (if $C = n-(k-q)$). Consequently, such a test is trivial in that all rejection probabilities (under the null as well as under the alternative) are zero (because of our Gaussianity assumption and the fact that all $\Sigma \in \mathfrak{C}_{Het}$ are positive definite). As an aside, we note that any $C < n-(k-q)$ leads to a nontrivial test as is easily seen.

Of course, a critical value C satisfying $C \geq n-(k-q)$ is certainly size-controlling, but is useless since it leads to a trivial test as just discussed. We now ask if and when the smallest size-controlling critical value $C_{\diamond}(\alpha)$, guaranteed to exist by Part (c) of Theorem 6.4, leads to a nontrivial test. (This is certainly so if α^* in Part (a) of Theorem 6.4 is positive, but note that the theorem is silent on this issue.) To obtain insight, we establish a simple, but important, relationship between the test statistics $\tilde{T}_{uc}$ and T_{uc} that is of independent interest also: note that standard least-squares theory gives

$$T_{uc}(y) = (n-k)\left(\sum_{i=1}^{n} \tilde{u}_i^2(y) / \sum_{i=1}^{n} \hat{u}_i^2(y) - 1\right)$$

for $y \notin \operatorname{span}(X)$, and recall $T_{uc}(y) = 0$ for $y \in \operatorname{span}(X)$. Hence, we obtain

$$\tilde{T}_{uc}(y) = (n-(k-q))\,(T_{uc}(y)/(n-k+T_{uc}(y)) = g(T_{uc}(y)) \qquad \textbf{(19)}$$

for every $y \notin \operatorname{span}(X)$, where $g : [0,\infty) \to [0, n-(k-q))$ is continuous and strictly increasing with $\lim_{x\to\infty} g(x) = (n-(k-q))$. (Since $T_{uc}(y_m) \to \infty$ for every sequence $y_m \to y \in \operatorname{span}(X) \backslash \mathfrak{M}_0$, the sharpness of the bound $n-(k-q)$ can thus also be read-off from (19).) As a consequence, for every critical value $C > 0$, the rejection regions $\{\tilde{T}_{uc} \geq C\}$ and $\{T_{uc} \geq g^{-1}(C)\}$ differ at most by $\operatorname{span}(X)$, which is a $\lambda_{\mathbb{R}^n}$-null set; in particular, the rejection probabilities (under the null as well as under the alternative) are the same.[47] *That is, the test statistics* $\tilde{T}_{uc}$ *and* T_{uc} *give rise to (essentially) the same test, if the critical values chosen are linked by the*

[46] If Assumption 1 is not satisfied, then $T_{Het} \equiv 0$, and the resulting test (with rejection region $\{T_{Het} \geq C\}$) is trivial as it never rejects for $C > 0$, while it always rejects for $C \leq 0$.

[47] This is so because of our Gaussianity assumption and the fact that all $\Sigma \in \mathfrak{C}_{Het}$ are positive definite.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

function g as above. In particular, as we shall see, this is the case if the respective smallest size-controlling critical values are used for both test statistics (provided both these values exist).

To see what the preceding discussion entails for the existence of nontrivial size-controlling critical values for $\tilde{T}_{uc}$, we distinguish two cases. In the first case, we shall see that nontrivial size-controlling critical values do not exist, whereas in the second case, they do indeed exist.

Case 1: Condition (8) is violated. Recall from Proposition 5.7 that then the size of $\{T_{uc} \geq D\}$ is 1 for every real D (in particular, implying that T_{uc} is not size-controllable). It transpires from the preceding discussion, that hence the size of $\{\tilde{T}_{uc} \geq C\}$ must equal 1 for every C satisfying $0 < C < n-(k-q)$ (and a fortiori for $C \leq 0$), because $D := g^{-1}(C)$ is well defined and real for $0 < C < n-(k-q)$. As a consequence, any size-controlling critical value C for $\tilde{T}_{uc}$ must satisfy $C \geq n-(k-q)$ (with the smallest size-controlling critical value given by $n-(k-q)$), thus leading to a rejection region that is trivial in that it is empty (if $C > n-(k-q)$) or is a $\lambda_{\mathbb{R}^n}$-null set, namely $\mathrm{span}(X)\backslash\mathfrak{M}_0$ (if $C = n-(k-q)$). That is—while $\tilde{T}_{uc}$ is size-controllable in the present case—it is so only in a trivial way.[48] (Another way of arriving at the above conclusion is to use Part (a) of Proposition 6.7 and to observe that in Part (a) of Theorem 6.4, the quantity C^* equals $n-(k-q)$. To see the latter, note that violation of condition (8) implies existence of an index $i \in I_1(\mathfrak{M}_0^{lin})$ with $e_i(n) \in \mathrm{span}(X)$. In particular, $\hat{u}(\mu_0 + e_i(n)) = 0$. Since $e_i(n) \notin \mathfrak{M}_0^{lin}$ must hold in view of $i \in I_1(\mathfrak{M}_0^{lin})$, and thus $\mu_0 + e_i(n) \notin \mathfrak{M}_0$ for every $\mu_0 \in \mathfrak{M}_0$ must be true, we may use (18) to arrive at $\tilde{T}_{uc}(\mu_0 + e_i(n)) = n-(k-q)$ for this $i \in I_1(\mathfrak{M}_0^{lin})$. This shows $C^* \geq n-(k-q)$. Equality then follows since $C^* \leq n-(k-q)$ trivially holds by (18). As a point of interest, we also note that $C^* = n-(k-q)$ implies that α^* in Part (a) of Theorem 6.4 satisfies $\alpha^* = 0$.)

Case 2: Condition (8) is satisfied. In this case T_{uc} is size-controllable according to Theorem 5.1. In particular, for any given $\alpha \in (0,1)$, there exists a smallest real number $D_\diamond(\alpha)$ such that the size of $\{T_{uc} \geq D_\diamond(\alpha)\}$ is less than or equal to α, with equality holding for $\alpha \in (0, \alpha^*_{T_{uc}}] \cap (0,1)$ where $\alpha^*_{T_{uc}}$ refers to α^* appearing in Theorem 5.1(a), and recall from that theorem that $\alpha^*_{T_{uc}} > 0$; and $D_\diamond(\alpha) > 0$ by Remark 5.4.[49] Also, note that the rejection region $\{T_{uc} \geq D_\diamond(\alpha)\}$ is not trivial as it has positive $\lambda_{\mathbb{R}^n}$-measure (and the same is true for its complement) (see the discussion at the very beginning of Section 6.2). Setting $C_\diamond(\alpha) = g(D_\diamond(\alpha))$ and using that $\{\tilde{T}_{uc} \geq C_\diamond(\alpha)\}$ and $\{T_{uc} \geq g^{-1}(C_\diamond(\alpha))\} = \{T_{uc} \geq D_\diamond(\alpha)\}$ differ at most by the $\lambda_{\mathbb{R}^n}$-null set $\mathrm{span}(X)$, we see that (i) $0 < C_\diamond(\alpha) < n-(k-q)$, (ii) the size of $\{\tilde{T}_{uc} \geq C_\diamond(\alpha)\}$ is less than or equal to α, with equality holding for $\alpha \in (0, \alpha^*_{T_{uc}}] \cap (0,1)$, (iii) $C_\diamond(\alpha)$ is the smallest size-controlling critical value (recall that g is strictly increasing), and (iv) the rejection region $\{\tilde{T}_{uc} \geq C_\diamond(\alpha)\}$ is not trivial as it has positive $\lambda_{\mathbb{R}^n}$-measure (and the same is true for its complement). In particular,

[48] The trivial size-controlling critical values C for $\tilde{T}_{uc}$ sort of correspond to using ∞ as a "size-controlling critical value" for T_{uc}.

[49] If $\alpha^*_{T_{uc}} < \alpha < 1$, then the size, in fact, equals $\alpha^*_{T_{uc}}$ (see Remark 5.6).

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

note that $\tilde{T}_{uc}$ and T_{uc} give rise to (essentially) the same test if the respective smallest size-controlling critical values are used. We furthermore note that in the present situation $C^*_{\tilde{T}_{uc}} = g(C^*_{T_{uc}})$ and $\alpha^*_{\tilde{T}_{uc}} = \alpha^*_{T_{uc}}$ hold, where $C^*_{T_{uc}}$, $\alpha^*_{T_{uc}}$ correspond to C^*, α^* in Part (a) of Theorem 5.1, whereas $C^*_{\tilde{T}_{uc}}$, $\alpha^*_{\tilde{T}_{uc}}$ correspond to C^*, α^* in Part (a) of Theorem 6.4.[50] In particular, $\alpha^*_{\tilde{T}_{uc}} > 0$ and $0 \leq C^*_{\tilde{T}_{uc}} < n-(k-q)$ follow. These claims can be seen as follows: under condition (8), we have $\mu_0 + e_i(n) \notin \operatorname{span}(X)$ for every $i \in I_1(\mathfrak{M}_0^{lin})$ and every $\mu_0 \in \mathfrak{M}_0$. Consequently, $\tilde{T}_{uc}(\mu_0+e_i(n)) = g(T_{uc}(\mu_0+e_i(n)))$, which proves $C^*_{\tilde{T}_{uc}} = g(C^*_{T_{uc}})$ in view of strict monotonicity of g. The relation $\alpha^*_{\tilde{T}_{uc}} = \alpha^*_{T_{uc}}$ then follows from the definitions of $\alpha^*_{\tilde{T}_{uc}}$ and $\alpha^*_{T_{uc}}$ using that $\{\tilde{T}_{uc} \geq C\}$ and $\{T_{uc} \geq g^{-1}(C)\}$ differ at most by the $\lambda_{\mathbb{R}^n}$-null set $\operatorname{span}(X)$ for every $C > 0$. Positivity of $\alpha^*_{\tilde{T}_{uc}}$ now follows from positivity of $\alpha^*_{T_{uc}}$ discussed before, and $C^*_{\tilde{T}_{uc}} < n-(k-q)$ follows since $C^*_{\tilde{T}_{uc}} = g(C^*_{T_{uc}})$ and $C^*_{T_{uc}} < \infty$. (Another way of proving $\alpha^*_{\tilde{T}_{uc}} > 0$ and $0 \leq C^*_{\tilde{T}_{uc}} < n-(k-q)$ without using relationship (19), is to first establish $C^*_{\tilde{T}_{uc}} < n-(k-q)$ [from observing that $\hat{u}(\mu_0+e_i(n)) \neq 0$ (as $\mu_0+e_i(n) \notin \operatorname{span}(X)$) for every $i \in I_1(\mathfrak{M}_0^{lin})$, which implies $\tilde{T}_{uc}(\mu_0+e_i(n)) < n-(k-q)$ for every such i in view of (18)] and then to proceed analogously as in the proof of Theorem 6.12.)

While $\tilde{T}_{uc}$ is always size-controllable, whereas T_{uc} is not, this does not represent any real advantage of $\tilde{T}_{uc}$ over T_{uc}, as we have seen that $\tilde{T}_{uc}$ admits only trivial size-controlling critical values in the case where T_{uc} is not size-controllable. Even more importantly, and already noted above, these test statistics give rise to (essentially) the same test if for both test statistics the respective smallest size-controlling critical values are used (provided they both exist).

6.2.2. *The Case of $\tilde{T}_{Het}$.* For $\tilde{T}_{Het}$, we find that, not infrequently, it is also a bounded function, although we have no proof that this is always so. We illustrate the problems that can arise here first by an example. See also Remark 6.14.

Example 6.1. Consider the $n \times 2$ design matrix X where the first column represents an intercept, the second column is $x := (1,-1,0,\ldots,0)'$, and $n \geq 3$. Let $R = (0,1)$, $r = 0$, and hence $q = 1$. Obviously, the first column of X spans $\mathfrak{M}_0^{lin}$. Since $e_i(n) \notin \mathfrak{M}_0^{lin}$, for every $i = 1,\ldots,n$, Assumption 2 holds. Furthermore, $\tilde{h}_{ii} = n^{-1}$. Thus, $\tilde{d}_i = \tilde{d}_1$ holds for every $i = 1,\ldots,n$ and for every of the five choices HC0R–HC4R. Note that $\tilde{d}_1^{-1} = 1$ (HC0R), $\tilde{d}_1^{-1} = 1-n^{-1}$ (HC1R), $\tilde{d}_1^{-1} = 1-n^{-1}$ (HC2R), $\tilde{d}_1^{-1} = (1-n^{-1})^2$ (HC3R), and $\tilde{d}_1^{-1} = 1-n^{-1}$ (HC4R), and hence $0 < \tilde{d}_1^{-1} \leq 1$ for all five choices. Straightforward computations now show that $\tilde{\Omega}_{Het}(y) = \tilde{d}_1\left[(y_1-\bar{y})^2+(y_2-\bar{y})^2\right]/4$ and

$$\tilde{T}_{Het}(y) = \tilde{d}_1^{-1}(y_1-y_2)^2 \big/ \left[(y_1-\bar{y})^2+(y_2-\bar{y})^2\right] \qquad \textbf{(20)}$$

[50] If $\alpha^*_{\tilde{T}_{uc}} < \alpha < 1$, then the size of $\{\tilde{T}_{uc} \geq C_\diamond(\alpha)\}$ is, in fact, equal to $\alpha^*_{T_{uc}} = \alpha^*_{\tilde{T}_{uc}}$ (cf. Footnote 49 and Remark 6.8).

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

whenever the numerator is positive, and $\tilde{T}_{Het}(y) = 0$ otherwise. Here, $\bar{y}$ denotes the arithmetic mean of the observations y_i. (For later use, we also note that the set $\tilde{\mathsf{B}}$ is given by $\{y \in \mathbb{R}^n : y_1 = y_2 = \bar{y}\}$, and that the size control condition (17) is satisfied, since $e_i(n) \notin \tilde{\mathsf{B}}$ for every $i = 1, \ldots, n$ [also note that μ_0 can be chosen to be zero because of $r = 0$]. Furthermore, $\tilde{T}_{Het}$ is not constant on $\mathbb{R}^n \backslash \tilde{\mathsf{B}}$, since $\tilde{T}_{Het}(e_1(n)) = \tilde{T}_{Het}(e_2(n)) = \tilde{d}_1^{-1} n^2/[(n-1)^2+1]$ and $\tilde{T}_{Het}(e_i(n)) = 0$ for $i \geq 3$ [note $n \geq 3$] and since $e_i(n) \notin \tilde{\mathsf{B}}$ for every i.) It is now evident from (20) that $\tilde{T}_{Het}(y) \leq 2\tilde{d}_1^{-1}$ for every $y \in \mathbb{R}^n$ and that this bound is attained whenever $y_1 + y_2 = 2\bar{y}$ and $y_1 \neq y_2$ (e.g., for $y = x$). It follows that any critical value $C \geq 2\tilde{d}_1^{-1}$ leads to a test with rejection region that is empty if $C > 2\tilde{d}_1^{-1}$, and is a Lebesgue null set if $C = 2\tilde{d}_1^{-1}$ (the latter following from Lemma D.1(d) in Appendix D of the Supplementary Material together with some of the observations just noted after (20)); thus, in both cases, all the rejection probabilities are zero under the null as well as under the alternative (given our Gaussianity assumption and the fact that all $\Sigma \in \mathfrak{C}_{Het}$ are positive definite); in particular, these tests have zero power. Since $\tilde{d}_1^{-1} \leq 1$, this eliminates all critical values $C \geq 2$ from practical use. In particular, this eliminates the commonly used choice where C is the 95% quantile of a chi-square distribution with 1 degree of freedom, which is approximately equal to 3.8415.

In the preceding example, any critical value $C \geq 2\tilde{d}_1^{-1}$ is trivially a size-controlling critical value for the given significance level α $(0 < \alpha < 1)$, but it is “too large” and leads to a trivial test. Certainly, one would prefer to use the smallest size-controlling critical value $C_\diamond(\alpha)$ instead (which in the preceding example exists by Theorem 6.4 and by what has been shown in the example) and one would hope that the resulting test is not trivial. As we shall show, this is indeed the case. To this end, we first give a general result that, in particular, is applicable to the preceding example. Recall that $C_\diamond(\alpha)$ is positive (Remark 6.6), and that Theorem 6.4 is silent on whether $\alpha^* > 0$ or not.

THEOREM 6.12. *Suppose Assumption 2 and (17) are satisfied and that $\tilde{T}_{Het}$ is not constant on $\mathbb{R}^n \backslash \tilde{\mathsf{B}}$. Let α satisfy $0 < \alpha < 1$, and let C^* and α^* be as defined in Part (b) of Theorem 6.4. If $C^* < \sup_{y \in \mathbb{R}^n} \tilde{T}_{Het}(y)$ holds, then we have $\alpha^* > 0$, and the rejection region $\{\tilde{T}_{Het} \geq C_\diamond(\alpha)\}$ is not a $\lambda_{\mathbb{R}^n}$-null set, where $C_\diamond(\alpha)$ is the smallest size-controlling critical value as in Part (c) of Theorem 6.4.*

Remark 6.13. (i) The preceding theorem clearly implies that—under its assumptions—the rejection probabilities associated with the rejection region $\{\tilde{T}_{Het} \geq C_\diamond(\alpha)\}$ are positive under the null as well as under the alternative (in view of our Gaussianity assumption and the fact that all $\Sigma \in \mathfrak{C}_{Het}$ are positive definite). (While we already know from Theorem 6.4(b) and Remark 6.8 that the rejection region $\{\tilde{T}_{Het} \geq C_\diamond(\alpha)\}$ has size equal to α in case $\alpha \in (0, \alpha^*] \cap (0,1)$, and has size equal to α^* if $\alpha^* < \alpha < 1$, this by itself does *not* allow one to conclude that the rejection region has positive $\lambda_{\mathbb{R}^n}$-measure as the case $\alpha^* = 0$ is not ruled out by Theorem 6.4(b) and Remark 6.8.)

(ii) Suppose $C^* = \sup_{y \in \mathbb{R}^n} \tilde{T}_{Het}(y)$, but that the other assumptions of Theorem 6.12 hold. [51] Then the rejection region $\{\tilde{T}_{Het} \geq C_\diamond(\alpha)\}$ is a $\lambda_{\mathbb{R}^n}$-null set; thus, also the smallest (and hence any) size-controlling critical value leads to a trivial test. To prove the claim, note that by Proposition 6.7 we have $C_\diamond(\alpha) \geq C^*$, implying that the rejection regions are either empty or coincide with the sets $\{\tilde{T}_{Het} = C^*\}$, respectively. In the latter case, apply Part (d) of Lemma D.1 in Appendix D of the Supplementary Material. We also point out that in the present case, $\alpha^* = 0$ must hold since the rejection regions appearing in the definition of α^* are all empty (because of $C > C^* = \sup_{y \in \mathbb{R}^n} \tilde{T}_{Het}(y)$ in the definition of α^*).

(iii) If Assumption 2 holds, but $\tilde{T}_{Het}$ is constant on $\mathbb{R}^n \backslash \tilde{\mathsf{B}}$, any rejection region of the form $\{\tilde{T}_{Het} \geq C\}$ is trivial in that the rejection region or its complement is a $\lambda_{\mathbb{R}^n}$-null set. (This case can actually occur [see Remark D.2 in Appendix D of the Supplementary Material].) If Assumption 2 is violated, $\tilde{T}_{Het}$ is identically zero and a similar comment applies.

Example 6.2. We continue the discussion of Example 6.1. As noted prior to Theorem 6.12, any critical value $C \geq 2\tilde{d}_1^{-1}$ is size-controlling in a trivial way, but leads to trivial rejection regions. We now show that the smallest size-controlling critical value $C_\diamond(\alpha)$ indeed leads to a nontrivial test (which, in particular, has positive rejection probabilities in view of our Gaussianity assumption and the fact that all $\Sigma \in \mathfrak{C}_{Het}$ are positive definite). For this, it suffices to verify the assumptions of Theorem 6.12. The first three assumptions have already been verified above. From the calculations in Example 6.1, it is now easy to see that $C^* = \tilde{d}_1^{-1} n^2 / [(n-1)^2 + 1]$, which is smaller than $2\tilde{d}_1^{-1} = \sup_{y \in \mathbb{R}^n} \tilde{T}_{Het}(y)$. This completes the proof of the assertion. From Remark 6.13(i), we furthermore see that the rejection region $\{\tilde{T}_{Het} \geq C_\diamond(\alpha)\}$ has size equal to α if $\alpha \in (0, \alpha^*] \cap (0,1)$, and has size equal to α^* if $\alpha^* < \alpha < 1$. Finally, we note that size-controlling critical values that do not lead to trivial tests must lie in the interval $[\tilde{d}_1^{-1} n^2 / [(n-1)^2 + 1], 2\tilde{d}_1^{-1})$ which is quite narrow as it is contained in the interval $[\tilde{d}_1^{-1}, 2\tilde{d}_1^{-1})$.

While the situation in Example 6.1 is somewhat particular, the example may perhaps contribute to a better understanding of the Monte Carlo findings in Davidson and MacKinnon (1985) and Godfrey (2006), namely that the tests, obtained from $\tilde{T}_{Het}$ (employing HC0R–HC4R weights) in conjunction with conventional critical values such as the 95% quantile of a chi-square distribution with appropriate degrees of freedom, can suffer from severe underrejection under the null.

Remark 6.14. Another class of examples where $\tilde{T}_{Het}$ is bounded is the case $q = k$ discussed in Remark 6.10. Recall from that remark that in case $q = k > 1$, condition (17) is, however, never satisfied and thus Theorem 6.12 is then not applicable. We have not further investigated nontriviality of tests based on $\tilde{T}_{Het}$ in

[51] We have not investigated whether this case can actually occur for $\tilde{T}_{Het}$. Recall that for $\tilde{T}_{uc}$ this case indeed can occur (see Case 1 in Section 6.2.1).

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

case $q = k$ beyond the observations made in Remark 6.10(iii) that constancy of $\tilde{T}_{Het}$ on $\mathbb{R}^n \backslash \tilde{\mathsf{B}}$ is possible in case $q = k \geq 1$ and thus then Remark 6.13(iii) applies.

7. GENERALIZATIONS

7.1. Generalizations Beyond Gaussianity

(i) All results in the preceding sections (as well as the extensions described in Appendix A of the Supplementary Material) referring to properties under the null hypothesis carry over as they stand to the situation where the error term $\mathbf{U}$ in (1) is elliptically symmetric distributed and has no atom at zero, i.e., $\mathbf{U}$ is distributed as $\sigma \Sigma^{1/2}\mathbf{z}$ where $\mathbf{z}$ has a spherically symmetric distribution on $\mathbb{R}^n$ that has no atom at zero.[52] This is so since—under this distributional model—the null rejection probabilities of any $G(\mathfrak{M}_0)$-invariant rejection region coincide with the corresponding null rejection probabilities under the Gaussian model (i.e., where $\mathbf{z}$ is standard Gaussian); see the discussion in Section 5.5 of Preinerstorfer and Pötscher (2016) and Appendix E.1 of Pötscher and Preinerstorfer (2018).[53] This implies, in particular, not only that the sufficient conditions for size controllability under the above elliptically symmetric distributed model as well as under the Gaussian model are the same, but also that the numerical values of the size-controlling critical values coincide. As a consequence, the algorithms for computing the size-controlling critical values in the Gaussian case (used in Section 11 and described in Section 10 and Appendix E of the Supplementary Material) can be used in the above elliptically symmetric distributed case without any change whatsoever. The same is actually true if $\mathbf{z}$ has a distribution in a certain class larger than the class of spherical symmetric distributions with no atom at zero (see Appendix E.1 of Pötscher and Preinerstorfer, 2018).

(ii) Furthermore, as discussed in detail in Appendix E.2 of Pötscher and Preinerstorfer (2018), the sufficient conditions for size controllability that we have derived under Gaussianity also imply size controllability for many more forms of distribution of $\mathbf{z}$ than those mentioned in (i); however, the corresponding size-controlling critical values may then differ from the size-controlling critical values that apply under Gaussianity.

(iii) Similarly as in Section 5.5 of Preinerstorfer and Pötscher (2016), the negative results given in the preceding sections (as well as the ones described in Appendix A of the Supplementary Material) such as e.g., size 1 results, extend in a trivial way beyond the Gaussian model as long as the maintained assumptions on the feasible error distributions are weak enough to ensure that the implied (possibly semiparametric) model, i.e., set of distributions for $\mathbf{Y}$, contains the set given in (2), but possibly contains also other distributions.

[52] Note that all results in the preceding sections (as well as the extensions in Appendix A of the Supplementary Material), except for a few comments in Section 6.2, are results referring to properties under the null hypothesis.

[53] Note that all rejection regions considered in the preceding sections are $G(\mathfrak{M}_0)$-invariant, because the test statistics considered are so.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

(iv) A further generalization beyond Gaussianity in the important special case where $\mathfrak{C} = \mathfrak{C}_{Het}$ is as follows: suppose $\mathbf{U}$ is distributed as $\sigma\Sigma^{1/2}\operatorname{diag}(\mathbf{r})\mathbf{z}$ where $\mathbf{z}$ is standard normally distributed on $\mathbb{R}^n$ and where the n-dimensional random vector $\mathbf{r}$ is independent of $\mathbf{z}$ with distribution ρ, where ρ is a distribution on $(0,\infty)^n$. (This includes the case where the elements of $\operatorname{diag}(\mathbf{r})\mathbf{z}$ form an i.i.d. sample from a scale mixture of normals.) Let $Q_{\mu,\sigma^2\Sigma,\rho}$ denote the implied distribution for $\mathbf{Y}$ given by (1) where $\mu = X\beta$. Consider now instead of (2) the (semiparametric) model given by all distributions $Q_{\mu,\sigma^2\Sigma,\rho}$ where $\mu \in \operatorname{span}(X)$, $0 < \sigma^2 < \infty$, $\Sigma \in \mathfrak{C}$, and ρ is an arbitrary distribution on $(0,\infty)^n$. Then the sufficient conditions for size controllability derived under Gaussianity in earlier sections (and in Appendix A of the Supplementary Material) also imply size controllability in this larger model. In fact, the size-controlling critical values that apply under Gaussianity deliver also size control under this more general model. This follows from the following reasoning: let W be a Borel set in $\mathbb{R}^n$ such that $P_{\mu_0,\sigma^2\Sigma}(W) \leq \alpha$ for every $\mu_0 \in \mathfrak{M}_0$, every $0 < \sigma^2 < \infty$, and every $\Sigma \in \mathfrak{C}_{Het}$. Then, for every such μ_0, σ^2, Σ, and every distribution ρ on $(0,\infty)^n$, we have

$$\begin{aligned} Q_{\mu_0,\sigma^2\Sigma,\rho}(W) &= \Pr(\mu_0 + \sigma\Sigma^{1/2}\operatorname{diag}(\mathbf{r})\mathbf{z} \in W) = \mathbb{E}[\Pr(\mu_0 + \sigma\Sigma^{1/2}\operatorname{diag}(\mathbf{r})\mathbf{z} \in W \mid \mathbf{r})] \\ &= \mathbb{E}[\Pr(\mu_0 + \sigma_{\mathbf{r}}\Sigma_{\mathbf{r}}^{1/2}\mathbf{z} \in W \mid \mathbf{r})] = \mathbb{E}[P_{\mu_0,\sigma_{\mathbf{r}}^2\Sigma_{\mathbf{r}}}(W)] \leq \alpha, \end{aligned}$$

where $\Sigma_{\mathbf{r}}^{1/2} := \Sigma^{1/2}\operatorname{diag}(\mathbf{r})/s_{\mathbf{r}}$ with $s_{\mathbf{r}}$ denoting the positive square root of the sum of the diagonal elements of $(\Sigma^{1/2}\operatorname{diag}(\mathbf{r}))^2 = \Sigma\operatorname{diag}^2(\mathbf{r})$ and where $\sigma_{\mathbf{r}} = \sigma s_{\mathbf{r}}$. Here, we have used that $P_{\mu,\sigma_{\mathbf{r}}^2\Sigma_{\mathbf{r}}}(W) \leq \alpha$ by assumption since $\Sigma_{\mathbf{r}} = \Sigma\operatorname{diag}^2(\mathbf{r})/s_{\mathbf{r}}^2 \in \mathfrak{C}_{Het}$ and $0 < \sigma_{\mathbf{r}} < \infty$ hold for every realization of $\mathbf{r}$. In the above, Pr denotes the probability measure governing $(\mathbf{r},\mathbf{z})$ and $\mathbb{E}$ the corresponding expectation operator. As a consequence, the smallest size-controlling critical value under Gaussianity is also the smallest size-controlling critical value under the semiparametric model considered here, as the latter model contains the Gaussian model as a submodel. (In the special case where $\operatorname{diag}(\mathbf{r})$ is a (random) multiple of the identity matrix I_n, the assumption $\mathfrak{C} = \mathfrak{C}_{Het}$ is superfluous as then $\Sigma_{\mathbf{r}} = \Sigma$, which by assumption belongs to the given $\mathfrak{C}$. In this case, $\mathbf{U}$ satisfies the assumptions in (i), and hence (iv) adds little new, except that—in contrast to (i)—the reasoning works without use of $G(\mathfrak{M}_0)$-invariance.)

(v) It is apparent from the reasoning in (iv) that Gaussianity of $\mathbf{z}$ can be replaced by any other distributional assumption for which size controllability has already been established. For example, one can in (iv) choose $\mathbf{z}$ to have a spherically symmetric distribution without an atom at zero or to have a distribution in the more general class mentioned in (i) (note that all relevant rejection regions discussed in earlier sections are $G(\mathfrak{M}_0)$-invariant and thus (i) applies). In a similar vein, one can combine the results in Appendix E.2 of Pötscher and Preinerstorfer (2018) discussed in (ii) above with the reasoning outlined in (iv). We abstain from presenting details.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

7.2. Generalizations to Stochastic Regressors

The assumption of nonstochastic regressors can be easily relaxed as follows: suppose X is random and $\mathbf{U}$ is conditionally on X distributed as $N(0,\sigma^2\Sigma)$, with $\sigma^2 = \sigma^2(X) > 0$ and $\Sigma = \Sigma(X) \in \mathfrak{C}_{Het}$ where $\sigma^2(\cdot)$ and $\Sigma(\cdot)$ may vary in given classes of functions. The size control results such as Theorems 5.1 and 6.4 can then obviously be applied after one conditions on X provided almost all realizations of X satisfy the assumptions of those theorems, which will typically be the case (for brevity, we do not provide a formal statement here).[54] The resulting conditional size control statements then immediately imply that the so-obtained conditional size-controlling critical values $C = C(\alpha, X)$ also control size unconditionally. Size 1 results such as, Propositions 5.5, 5.7, or 6.7, also extend to conditional size 1 results in a similar manner provided $\sigma^2(X)$ and $\Sigma(X)$ vary independently through all of $(0,\infty)$ and $\mathfrak{C}_{Het}$, respectively, for (almost) every realization of X, when the functions $\sigma^2(\cdot)$ and $\Sigma(\cdot)$ vary in the before-mentioned function classes.[55] Generalizations to non-Gaussianity similarly as discussed in Section 7.1 are also possible in the present context.

8. RESULTS FOR OTHER CLASSES OF TESTS

The results in Sections 5 and 6 (and in Appendix A of the Supplementary Material) have been obtained with the help of a general theory developed in Section 5 of Preinerstorfer and Pötscher (2016), Section 5 of Pötscher and Preinerstorfer (2018), and Section 3.1 of Pötscher and Preinerstorfer (2019) that covers a very broad class of test statistics (and actually allows also for correlated errors). We note that, like in Section 7.1, Gaussianity is again not essential for a good portion of this general theory (see Section 5.5 of Preinerstorfer and Pötscher, 2016 as well as Appendix E of Pötscher and Preinerstorfer, 2018).[56] We next discuss a few further situations that can also be handled by the general theory just mentioned, but we refrain from spelling out the details:[57]

(i) The test statistic considered is an OLS-based test statistic like T_{Het}, but where $\hat{\Omega}_{Het}$ is now replaced by an appropriate estimator derived from a given (possibly misspecified) *parametric* heteroskedasticity model described by a parameter vector θ.

(ii) The test statistic is a Wald-type test statistic based on a (feasible) generalized least-squares estimator together with an appropriate covariance matrix estimator based on a given (possibly misspecified) parametric model. (This includes the

[54] An appropriately modified statement applies to the size control results in Appendix A of the Supplementary Material.

[55] See Footnote 40 in Pötscher and Preinerstorfer (2023) for a discussion of sufficient conditions.

[56] Also, arguments like in (iv) and (v) of Section 7.1 can be applied to try to obtain generalizations.

[57] Applying some of the main results of this general theory (e.g., Corollary 5.6 or Proposition 5.12 of Pötscher and Preinerstorfer, 2018) will require one to determine the set $\mathbb{J}(\mathcal{L},\mathfrak{C})$ defined in Appendix B of the Supplementary Material. For the important cases $\mathfrak{C} = \mathfrak{C}_{Het}$ and $\mathfrak{C} = \mathfrak{C}_{(n_1,\ldots,n_m)}$ (defined in Appendix A of the Supplementary Material), this is already accomplished in Propositions B.1 and B.2 in Appendix B of the Supplementary Material.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

(quasi-)maximum likelihood estimator (provided θ is unrelated to β).) Alternatively, the test statistic is the (quasi-)likelihood ratio or (quasi-)score test statistic based on this parametric model.

(iii) The test statistic is a Wald-type test statistic as in (ii), except that the covariance matrix estimator is now nonparametric (in the spirit of heteroskedasticity robust testing) as described in Romano and Wolf (2017). See also Cragg (1983, 1992), Flachaire (2005), Wooldridge (2010, 2012), Romano and Wolf (2017), Lin and Chou (2018), and DiCiccio, Romano, and Wolf (2019).

9. SOME COMMENTS ON POWER

Under our maintained assumptions, heteroskedasticity robust tests based on T_{Het} or T_{uc} (using an arbitrary critical value C, including size-controlling ones) have positive power everywhere in the alternative (cf. the discussion at the beginning of Section 6.2). These tests can furthermore be shown to have power that goes to one as one moves away from the null hypothesis along sequences $(\mu_l, \sigma_l^2, \Sigma_l)$ where μ_l moves further and further away from $\mathfrak{M}_0$ (the affine space of means described by the restrictions $R\beta = r$) in an orthogonal direction as $l \to \infty$, where σ_l^2 converges to some finite and positive σ^2, and Σ_l converges to a *positive definite* matrix. Despite of what has just been said, these tests can have, in fact not infrequently will have, *infimal* power equal to zero if $\mathfrak{C}$ is sufficiently rich, e.g., if $\mathfrak{C} = \mathfrak{C}_{Het}$ (cf. Theorem 4.2 in Preinerstorfer and Pötscher, 2016, Lemma 5.11 in Pötscher and Preinerstorfer, 2018, and Theorem 4.2 in Pötscher and Preinerstorfer, 2019). (This does not contradict the before-mentioned result as for this result sequences Σ_l that converge to a singular matrix as $l \to \infty$ were ruled out.)

For tests based on $\tilde{T}_{Het}$ or $\tilde{T}_{uc}$, the situation is somewhat different. As shown in Section 6.2, tests based on $\tilde{T}_{Het}$ or $\tilde{T}_{uc}$ can be trivial for some choices of critical values C (and then will have power zero everywhere in the alternative). However, if C is chosen to be the smallest size-controlling critical value (provided it exists), the resulting tests obtained from $\tilde{T}_{Het}$ or $\tilde{T}_{uc}$ will typically have positive power (under appropriate assumptions). In particular, then the test based on $\tilde{T}_{uc}$ has the same power function as the test based on T_{uc} that uses its smallest size-controlling critical value, provided the latter exists (see Section 6.2.1). We have not further investigated the power properties of the tests based on $\tilde{T}_{Het}$ in any more detail on a theoretical level. The numerical results in Section 11.2 seem to suggest that for these tests, power may not go to one along sequences $(\mu_l, \sigma_l^2, \Sigma_l)$ as mentioned above: in fact, power does not rise above the significance level α in some examples (on the range of alternatives considered). This feature makes tests based on $\tilde{T}_{Het}$ rather undesirable.

10. COMPUTING THE SIZE AND SMALLEST SIZE-CONTROLLING CRITICAL VALUES

Consider a testing problem as in equation (3) with $\mathfrak{C} = \mathfrak{C}_{Het}$, and let T be one of the test statistics considered in the present article (e.g., T_{Het} with some choice for the weights d_i). Suppose we want to numerically determine the size of the test with

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

rejection region $\{T \geq C\}$ for some user-supplied critical value C, i.e., we want to determine

$$\sup_{\mu_0 \in \mathfrak{M}_0} \sup_{0<\sigma^2<\infty} \sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0,\sigma^2\Sigma}(T \geq C). \tag{21}$$

Now, for all test statistics T considered in the present article, this can be simplified to

$$\sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0,\Sigma}(T \geq C), \tag{22}$$

where, subject to $\mu_0 \in \mathfrak{M}_0$, μ_0 can be chosen as desired. This is due to invariance properties of T (cf. Remarks 3.2 and 6.2). The quantity in (22) can now be approximated numerically by any maximization algorithm where the probabilities are evaluated by Monte Carlo methods or by the algorithm described in Davies (1980) in case $q = 1$ (cf. Appendix E.1 of the Supplementary Material).[58]

Suppose next that we want to numerically determine the smallest size-controlling critical value $C_\diamond(\alpha) \in \mathbb{R}$ $(0 < \alpha < 1)$ when using the test statistic T. (We assume here that the user knows that the smallest size-controlling critical value indeed exists, e.g., because the user has checked that the sufficient conditions developed in the present article hold, or because of other reasoning as, e.g., used in Example 5.5.) Then, in view of (21) and (22), we need to compute $C_\diamond(\alpha)$ as the smallest real number C for which

$$\sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0,\Sigma}(T \geq C) \leq \alpha \tag{23}$$

holds. The quantity to the left in (23) is non-increasing in the critical value C. Hence, to determine the smallest size-controlling critical value $C_\diamond(\alpha)$, any line-search algorithm (in combination with an algorithm to determine the sizes as described before) can be used to compute $C_\diamond(\alpha)$. We stress that it is of foremost importance to know that the testing problem at hand actually allows for size control before one attempts to numerically determine $C_\diamond(\alpha)$. Hence, the theoretical results of the present article are of paramount importance also for the algorithmic aspect of the problem.

The specific algorithms we use to determine size and size-controlling critical values in our numerical studies are based on the above observations and are described in detail in Appendix E of the Supplementary Material. They are made available in the R package **hrt** (Preinerstorfer, 2021) for the convenience of the user. The numerical procedures we use are heuristic in nature. Questions of efficacy of these algorithms or about theoretical guarantees are certainly important, but are beyond the scope of the present article.

Determining smallest size-controlling values numerically is important, e.g., if one wants to compare their magnitude with that of standard critical values in some

[58] Alternative to Davies (1980), other algorithms like Imhof's algorithm, etc., can be used, some of which are also implemented in the R package **CompQuadForm** (Duchesne and de Micheaux, 2010).

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

special cases, as we do inter alia in the next section, or if one wants to obtain a confidence interval. However, a user who has observed the data and only wants to decide whether or not to reject the null hypothesis at significance level α $(0 < \alpha < 1)$ when using T combined with the smallest size-controlling critical value $C_{\diamond}(\alpha)$, can actually perform this test without needing to compute $C_{\diamond}(\alpha)$: let y_{obs} be the observed data. Define the "maximal p-value" as

$$\begin{aligned} p(y_{obs}) &= \sup_{\mu_0 \in \mathfrak{M}_0} \sup_{0<\sigma^2<\infty} \sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0, \sigma^2 \Sigma}(\{z \in \mathbb{R}^n : T(z) \geq T(y_{obs})\}) \\ &= \sup_{\Sigma \in \mathfrak{C}_{Het}} P_{\mu_0, \Sigma}(\{z \in \mathbb{R}^n : T(z) \geq T(y_{obs})\}), \end{aligned} \quad \textbf{(24)}$$

where the second equality in the display follows from the invariance properties mentioned before (and $\mu_0 \in \mathfrak{M}_0$ can be chosen as desired). It is now not difficult to see that $p(y_{obs}) \leq \alpha$ is equivalent to $T(y_{obs}) \geq C_{\diamond}(\alpha)$. That is, rejecting if and only if $p(y_{obs}) \leq \alpha$ leads to exactly the same test as rejecting if and only if $T(y_{obs}) \geq C_{\diamond}(\alpha)$, with the former description having the advantage that the more costly computation of $C_{\diamond}(\alpha)$ can be avoided. What needs to be computed is (24), which, however, is nothing else than the size of the test when using the "critical value" $T(y_{obs})$. Hence, $p(y_{obs})$ can be determined by any algorithm that determines the size (22) for the user-supplied "critical value" $C = T(y_{obs})$. In particular, the routine "size" provided in the R package **hrt** (Preinerstorfer, 2021) can be used for this purpose. Note that checking whether $p(y_{obs}) \leq \alpha$ avoids the line-search part (as outlined following (23)), and is thus computationally *more efficient* than first determining $C_{\diamond}(\alpha)$ (as outlined above) and then checking whether $T(y_{obs}) \geq C_{\diamond}(\alpha)$.

Finally, we note that if (contrary to what we assume in this section) no size-controlling critical value exists for a given significance level $\alpha \in (0,1)$, then the maximal p-value in (24) is larger than α for *every* possible observed value y_{obs}, and the corresponding test thus never rejects and thus is uninformative. Hence, while the explicit computation of a smallest size-controlling critical value can be avoided for performing a single test, knowing its existence is important as then the resulting test is guaranteed to be informative (nontrivial) if T_{Het} or T_{uc} is being used, and the same is true for $\tilde{T}_{Het}$ and $\tilde{T}_{uc}$ under the conditions discussed in Section 6.2.

We also note that in view of the discussion in Section 7.1, the algorithms for computing null rejection probabilities, size, and smallest size-controlling critical values discussed in this section and Appendix E of the Supplementary Material remain valid for elliptically symmetric distributed data without any need for modification. With regard to computing size and smallest size-controlling critical values, the same is also true for the semiparametric model described in (iv) of Section 7.1.

11. NUMERICAL RESULTS

In this section, we pursue two goals:

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

1. In Section 11.1, we show numerically that any of the usual heteroskedasticity robust tests can suffer from overrejection of the null hypothesis (sometimes by a large margin) when they are based on conventional critical values. While this adds to similar evidence already present in the literature for the HC0–HC4-based tests (see Section 1), this seems to be a new observation for the HC0R–HC4R-based tests. In any case, this drives home the point that none of these heteroskedasticity robust tests based on conventional critical values comes with a guarantee that size is controlled by the nominal significance level α. Consequently, instead of using conventional critical values, this strongly suggests to use (smallest) size-controlling critical values as investigated in this article.
2. In Section 11.2, we then numerically compute smallest size-controlling critical values and study the power behavior of tests based on such size-controlling critical values in some examples.

In this section (and in the attending Appendixes E and F of the Supplementary Material), we shall often refer to T_{Het} as HC0–HC4 when we want to stress that the weights d_i being used are the HC0–HC4 weights, respectively (see Section 3). Similarly, we shall refer to $\tilde{T}_{Het}$ as HC0R–HC4R when the HC0R–HC4R weights are used (see Section 6). For reasons of uniformity of notation, we shall then often denote T_{uc} as UC and $\tilde{T}_{uc}$ as UCR. Furthermore, throughout this section, we consider the heteroskedastic Gaussian linear model with $\mathfrak{C} = \mathfrak{C}_{Het}$ as introduced in Section 2; in particular, the notion of size in the present section (and the attending appendixes) *always* refers to this model.

The algorithms for computing rejection probabilities, the size of a test, and size-controlling critical values used in the before-mentioned numerical computations are described in Section 10 and Appendix E of the Supplementary Material. Implementations are available as an R package **hrt** (Preinerstorfer, 2021).

11.1. Tests Based on Conventional Critical Values

We consider the important case $q = 1$, and first illustrate numerically that none of the test statistics UC, HC0–HC4, UCR, and HC0R–HC4R combined with the critical value $C_{\chi^2,0.05} \approx 3.8415$ results in a test that is *guaranteed* to have size less than or equal to $\alpha = 0.05$. This is achieved by providing instances of design matrices X and of hypotheses, described by (R, r), such that the respective test has size larger than the nominal significance level $\alpha = 0.05$, often by a large margin. Here, $C_{\chi^2,0.05}$ denotes the 95%-quantile of a chi-square distribution with 1 degree of freedom. (This critical value has a justification for use with HC0–HC4 or HC0R–HC4R via asymptotic considerations, but, in general, there is no such justification for use with UC or UCR, which we nevertheless include here for completeness.[59]) That is, in the instances we exhibit, this conventional critical

[59] Of course, in the special case of homoskedasticity, the before-mentioned justification also applies to UC and UCR.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

value turns out to be *too small*. We next show similar results for other suggestions of critical values, e.g., for "degree-of-freedom" adjustments to the conventional chi-square-based critical value such as the Bell–McCaffrey adjustment (Bell and McCaffrey, 2002; Imbens and Kolesár, 2016). It is important to note here that in all the instances mentioned, our conditions for size controllability are satisfied, showing that size-controlling critical values can actually be found; hence, the overrejection problems mentioned before are *not* intrinsic problems, but only reflect the fact that conventional critical values can be a bad choice and do not guarantee size control. (In the present context, it is worth recalling that for the test statistics HC0R–HC4R, we have already shown in Example 6.1 in Section 6.2 that other situations can be found in which conventional critical values such as $C_{\chi^2,0.05}$ are *too large*, as the resulting tests reject with probability zero only (under the null as well as under the alternative), rendering these tests useless.)

To uncover instances where the conventional critical value $C_{\chi^2,0.05}$ is too small, we make use of the following observation: in case a given test statistic from the above list (together with a given design matrix X and hypothesis described by (R,r)) is such that the lower bound C^* on size-controlling critical values obtained in Proposition 5.5 (Proposition 6.7, respectively) exceeds $C_{\chi^2,0.05}$, we are done, as we then know that the critical value $C_{\chi^2,0.05}$ leads to a test that has size 1. (As noted subsequent to Theorems 5.1 and 6.4, the value of r actually plays no role here, and we may set it to zero.)

Since the lower bounds C^* for size-controlling critical values in Proposition 5.5 (Proposition 6.7, respectively) depend on the given test statistic, on X and on R, we may—for any given choice of test statistic and any given R—numerically search for particularly "hostile" design matrices, i.e., for design matrices for which the lower bound is large, to see whether matrices X exist for which the lower bound exceeds $C_{\chi^2,0.05}$. We only do this for $k = 2$, $R = (0,1)$, $r = 0$, and $n = 25$, and restrict ourselves to matrices X with first column representing an intercept. The concrete search used is detailed in Appendix F.1 of the Supplementary Material (see Algorithm 5 in particular). Table 1 provides, for every test statistic considered, the lower bound C^* corresponding to the most "hostile" design matrix found by the search. (As the searches are run separately for each test statistic, the resulting "hostile" design matrices will typically differ across the runs.)[60]

In combination with the theoretical results from Propositions 5.5 and 6.7, Table 1 shows that for some design matrices X, the critical value $C_{\chi^2,0.05} \approx 3.8415$ results in a test with size equal to 1 when combined with UC, HC0–HC2, and also with UCR. (This is so despite the fact that, for any of the 12 test statistics considered,

[60] Since, for example, HC0 is a multiple of HC1, where the factor is $n/(n-k) = 1.09$, we know that the "hostile" design matrix obtained from the search for HC1 leads to a C^*-value of $1.09 \times 1,711.19 = 1,865.20$ for HC0, larger than the value 95.56 obtained from the search for HC0 (cf. Table 1). We could have reported this larger value, but decided to present the raw results from our searches as this is sufficient for our purposes. We also note that our search procedure detailed in Appendix F.1 of the Supplementary Material does not seriously attempt to optimize the C^*-value (for every one of the test statistics considered) over the set of all feasible X, but is only a crude search for finding a matrix resulting in a C^*-value sufficiently large for our purposes.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

Table 1. C^* under respective "hostile" X.

UC	731.60	UCR	23.59
HC0	95.56	HC0R	1.08
HC1	1711.19	HC1R	1.04
HC2	52.23	HC2R	1.04
HC3	1.00	HC3R	1.00
HC4	1.02	HC4R	1.04

Table 2. "Worst-case" sizes using $C_{\chi^2,0.05}$.

UC	0.98	UCR	0.98
HC0	0.99	HC0R	0.16
HC1	1.00	HC1R	0.17
HC2	0.99	HC2R	0.17
HC3	0.19	HC3R	0.14
HC4	0.11	HC4R	0.10

the sufficient conditions for size control in the pertaining theorems in Sections 5 and 6.1 are satisfied for all relevant X matrices encountered in the numerical procedure [as we have checked], and hence it is known that size-controlling critical values exist in all these situations!) Table 1 is not informative about the size of the remaining seven tests, since the corresponding entries in that table are all less than $C_{\chi^2,0.05}$. To obtain insight into the sizes of the remaining seven tests, we do the following: for each of the tests, we numerically compute the size for various instances of design matrices (the ones that give rise to Table 1) and report the largest one of these sizes ("worst-case" sizes) in Table 2.[61] We actually do this for all the 12 tests considered. The algorithm used in the size computation is the implementation of Algorithm 1 in the R package **hrt** (Preinerstorfer, 2021) (cf. the description in Appendixes E.2 and F.1 of the Supplementary Material). Table 2 now clearly shows that for *every* test statistic considered, an instance can be found, in which the size of the test (when using the critical value $C_{\chi^2,0.05}$) clearly exceeds the nominal significance level $\alpha = .05$. The lowest value in that table is attained by HC4R, but a size of 0.10 is still twice the nominal significance level α.

We note that the numbers shown in Table 2 actually only represent numerically determined lower bounds for the actual sizes, as their computation involves (for any given X) a numerical search procedure (over the set $\mathfrak{C}_{Het}$) for the worst-case null rejection probability; that is, the numbers shown in Table 2 correspond to the null rejection probability computed from a "bad" covariance matrix Σ, but potentially not for the "worst" possible one. (In this process, for any given $\Sigma \in \mathfrak{C}_{Het}$, we have to numerically compute the null rejection probability, which can be done quite

[61] Of course, considering additional design matrices X would potentially lead to even larger sizes.

Table 3. "Worst-case" sizes using F-critical value.

UC	0.98	UCR	0.98
HC0	0.99	HC0R	0.15
HC1	1.00	HC1R	0.16
HC2	0.98	HC2R	0.15
HC3	0.18	HC3R	0.13
HC4	0.09	HC4R	0.08

accurately in case $q = 1$ by algorithms like the Davies algorithm [see Appendix E.1 as well as Appendix E.2 of the Supplementary Material].) In particular, the entries in the 0.98–0.99 range in Table 2 are numerically determined lower bounds for the size, which, in fact, we know to be equal to 1 in light of Table 1. (We could have used this knowledge to replace the entries in question in Table 2 by 1, but we decided otherwise in order to showcase the concrete outcome of the numerical algorithm that has been run. Of course, one could also improve this outcome by using a higher accuracy parameter in the optimization procedures involved.)

Sometimes—without much theoretical justification in general—it is suggested in the literature to replace $C_{\chi^2,0.05}$ by the 95% quantile of an $F_{1,n-k}$-distribution, which is approximately 4.28 in the situation considered here ($n-k=23$). Obviously, from Table 1, we see that the conclusions regarding UC, HC0–HC2, and UCR remain the same when this critical value is used. Repeating the exercise that has led to Table 2, but with $C_{\chi^2,0.05}$ replaced by the 95% quantile of an $F_{1,n-k}$-distribution, gives Table 3, leading essentially to the same conclusions.

"Degree-of-freedom" adjustments to the conventional chi-square-based critical value such as the Bell–McCaffrey adjustment (Bell and McCaffrey, 2002) have been discussed in the literature. In particular, Imbens and Kolesár (2016) suggested to use this adjustment with the HC2 statistic. We have repeated the above exercise that has led to the entry for HC2 in Table 2, but with $C_{\chi^2,0.05}$ replaced by the Bell–McCaffrey adjustment. For the computation of the Bell–McCaffrey adjustment, we relied on the R package **dfadjust** (Kolesár, 2019). For the resulting test, the largest size that was found in our computations was 0.24, which is more than four times the nominal significance level. It transpires that this adjustment does also not come with a size guarantee.

We conclude here by stressing that the negative findings in this subsection were obtained in a very simple model with only two regressors and where only one of the parameters is subject to test. For more complex models and test problems, the size distortions may even be worse.

11.2. Power Comparison of Tests Based on Size-Controlling Critical Values

A power comparison of two tests, both conducted at a given *nominal* significance level α, makes sense only if both tests actually are level α tests, i.e., if both tests have a size not exceeding the given α. For this reason, we now compare the

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

tests obtained from the statistics UC, HC0–HC4, UCR, and HC0R–HC4R only when respective *smallest size-controlling critical values* are used. Our theoretical results concerning the existence of size-controlling critical values, together with the algorithms for their computation in Appendix E of the Supplementary Material, allow for such a comparison in terms of power. In all cases considered in this section, $q = 1$ will hold.

Throughout, in addition to the power functions of the before-mentioned tests, we also show as a benchmark the power function of the *infeasible* (i.e., oracle) GLS-based F-test conducted at the 5% significance level, that makes use of knowledge of Σ. For given $\Sigma \in \mathfrak{C}_{Het}$, the distribution of this infeasible GLS-based F-test statistic is (under $P_{X\beta,\sigma^2\Sigma}$ with $\beta \in \mathbb{R}^k$, $\sigma^2 \in (0,\infty)$) a noncentral $F_{1,n-k}$-distribution with noncentrality parameter δ^2, where

$$\delta = (R(X'\Sigma^{-1}X)^{-1}R')^{-1/2}(R\beta - r)/\sigma.$$

Since the power functions of all the tests considered in our study depend on the parameters β, σ^2, and Σ only through $(R\beta - r)/\sigma$ and Σ (because of $G(\mathfrak{M}_0)$-invariance and Proposition 5.4 in Preinerstorfer and Pötscher, 2016), and thus depend only on δ and Σ, we shall—for given Σ—present all these power functions as a function of δ. We show only results for $\delta \geq 0$, as the power functions in fact depend on δ only through $|\delta|$ (for given Σ) (see Proposition 5.4 in Preinerstorfer and Pötscher, 2016).

11.2.1. *Comparing the Means of Two Heteroskedastic Groups.*

As a practically relevant example, we here compare the power of tests based on size-controlling critical values in the context of Example 5.4. That is, we treat the problem of comparing the means of two heteroskedastic groups (e.g., a treatment and a control group), the null hypothesis being that the difference of expected outcomes in each group is zero. We consider the case where $n = 30$ and $\alpha = 0.05$. Furthermore, we vary the size n_1 of the first group ($n_1 \in \{3, 9, 15\}$), corresponding to a “strongly unbalanced,” “moderately unbalanced,” and “balanced” design, respectively. We compute the power for a number of covariance matrices Σ_a given as follows: for $a = 1, 5, 9$, define

$$\Sigma_a = 10^{-1}\operatorname{diag}\left(\frac{a}{n_1}, \ldots, \frac{a}{n_1}, \frac{10-a}{n-n_1}, \ldots, \frac{10-a}{n-n_1}\right) \in \mathfrak{C}_{Het},$$

where the first n_1 (and the last $n - n_1$, respectively) diagonal entries of each Σ_a are constant. That is, we look at power functions evaluated at covariance matrices under which the subjects in the same group actually have the same variances. (For brevity, we do not report power functions for covariance matrices not sharing this property.) For the balanced design, we note that Σ_1 and Σ_9 lead to the same power of each test (but we report all results for completeness), and that Σ_5 corresponds to homoskedasticity.

The critical values are chosen in each case as the smallest critical value guaranteeing size control over $\mathfrak{C}_{Het}$ (implying, of course, that the corresponding

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

tests can have null rejection probabilities smaller than α for the covariance matrices Σ_a considered). The existence of said critical values follows from our theory and is discussed in detail in Example 5.4 for the test statistics UC and HC0–HC4; in particular, all assumptions of Theorems 5.1 are satisfied. For UCR, the existence is guaranteed by Part (a) of Theorem 6.4. With regard to the test statistics HC0R–HC4R, note that Assumption 2 is satisfied since $e_i(n) \notin \mathfrak{M}_0^{lin} = \mathfrak{M}_0 = \operatorname{span}((1,\ldots,1)')$ for every $i = 1,\ldots,n$ as $n = 30 > k = 2$. This also shows that the sufficient condition for size control (17) is satisfied as $\tilde{\mathsf{B}} = \mathfrak{M}_0$ is easily verified and since one may set $\mu_0 = 0$. We have verified the non-constancy assumption on the test statistics HC0R–HC4R in Theorem 6.4 numerically. As a consequence, all assumptions of Part (b) of Theorem 6.4 are satisfied.

We note that some of the test statistics differ from each other only by a known multiplicative constant and hence are equivalent in the sense that they give rise to the same test *when the respective smallest size-controlling critical value is employed* (see Remarks 5.3 and 6.5): in the unbalanced case ($n_1 \in \{3,9\}$), HC0 and HC1 are equivalent in this sense, as are HC0R–HC4R (the latter is so since $\tilde{h}_{ii} = 1/n$ which does not depend on i). In the balanced case ($n = 15$), UC and HC0–HC4 are all equivalent, and the same is true for UCR and HC0R–HC4R as is not difficult to see. Furthermore, in the balanced case as well as in the unbalanced case, the rejection regions of the tests based on UC and UCR coincide essentially (i.e., up to a $\lambda_{\mathbb{R}^n}$ -null set) as a consequence of the relationship established in Section 6.2.1. In particular, it follows that in the balanced case, *all* tests considered (essentially) coincide. We nevertheless compute the power functions for each of the tests separately without making use of the noted equivalencies; this provides a double check of our numerical results.[62]

Numerically, the critical values were determined through the implementation of Algorithms 1 and 3 in the R package **hrt** (Preinerstorfer, 2021) version 1.0.0, and the power functions were computed with the implementation of the algorithm by Davies (1980) in the R package **CompQuadForm** (Duchesne and de Micheaux, 2010) version 1.4.3 (see Appendixes E.2 and F.2 of the Supplementary Material for more details). For the sake of illustration, we also report the critical values obtained for every test considered and every balancedness condition in Table 4.

In relation to Table 4, we note that the equivalences discussed before predict, e.g., that the ratio between the entries in the column labeled HC0 and the corresponding entries in the column labeled HC1 should be equal to $n/(n-2) = 30/28 \approx 1.0714$. The ratios computed from the table are 1.0414, 1.0761, and 1.0721 (for $n_1 = 3,9,15$), which is in pretty good agreement (especially if one converts the critical values shown in the table to critical values for the corresponding "t -test" versions by computing their square roots). The agreement between theoretical and observed ratios for the HC0R–HC4R columns is similar. In the balanced case, one can also use the additional equivalences mentioned before

[62] The equivalencies mentioned in this paragraph for the two-group-comparison problem analogously hold for general n, n_1, and n_2 as is easily seen.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

Table 4. The smallest size-controlling critical values for comparing the means of two heteroskedastic groups.

n_1	UC	HC0	HC1	HC2	HC3	HC4
3	225.97	26.69	25.63	17.48	11.86	5.43
9	12.70	5.80	5.39	5.10	4.55	4.70
15	4.59	4.91	4.58	4.59	4.28	4.58

n_1	UCR	HC0R	HC1R	HC2R	HC3R	HC4R
3	25.82	3.25	3.14	3.14	3.02	3.13
9	9.05	4.28	4.15	4.14	4.06	4.19
15	4.08	4.23	3.92	4.08	3.95	4.09

and one again finds very good agreement. Similarly, the critical values for UC and UCR in Table 4 are in excellent agreement with their theoretical relationship found in Section 6.2.1. The reason for the small discrepancies observed lies in the fact that the algorithm underlying the computations for Table 4 makes use of a random search algorithm. Concerning Table 4, we also mention that, in the example considered here and for the test statistic HC2, Ibragimov and Müller (2016) prove in their Theorem 1 (see also the discussion preceding that theorem) that the smallest size-controlling critical values are given by 18.51 ($n_1 = 3$), 5.32 ($n_1 = 9$), and 4.60 ($n_1 = 15$), respectively. The numerically determined critical values in Table 4 are reasonably close to these values (after conversion of the critical values to corresponding "t-test" critical values the maximal difference is about 0.1). Of course, the accuracy of our algorithm could be increased by using more stringent accuracy parameters in the optimization routines underlying the computation of the critical value, but this would come with a longer runtime.

From Table 4, it is clear that for the tests based on *unrestricted* residuals, the smallest size-controlling critical values obtained are always larger, sometimes considerably, than $C_{\chi^2,0.05} \approx 3.8415$, again showing that the latter critical value is not effecting size control. For the tests based on *restricted* residuals, the smallest size-controlling critical values sometimes fall below $C_{\chi^2,0.05}$ in the strongly unbalanced case (which is not completely surprising in view of Section 6.2.2); while in this case $C_{\chi^2,0.05}$ effects size control, using the smaller size-controlling critical values given in Table 4 can only be advantageous in terms of power.

That being said, we emphasize a trivial, but important point, namely that comparing the magnitudes of size-controlling critical values relating to *different* test statistics is not very meaningful and, in particular, *not* a valid way of comparing the quality of the resulting tests. That is, while it may be tempting to infer from Table 4 that the HC0 test should be considerably more conservative than the HC4 test, or that the UC test should be considerably more conservative than the UCR test, such a conclusion would be false and not warranted at all (in particular, recall that UC and UCR in fact result in (essentially) the same test if the critical values

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

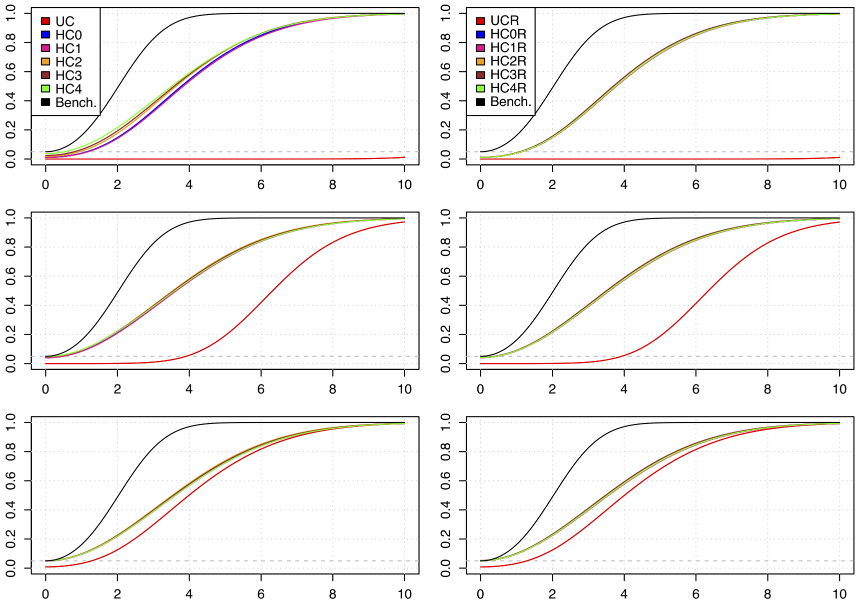

FIGURE 1. Power functions for $n_1 = 3$. Left column: tests based on unrestricted residuals (cf. legend). Right column: tests based on restricted residuals (cf. legend). The rows correspond to Σ_a for $a = 1, 5, 9$ from top to bottom. The abscissa shows δ. In the left panel, the HC0–HC4 curves turn out to be barely distinguishable, with the HC1 curve lying on top of the HC0 curve. In the right panel, the HC4R curve lies on top of the HC0R–HC3R curves. See the text for an explanation.

from Table 4 are being used). While this would be correct if the critical values were all meant to be used with the same test statistic (which they are not), critical values belonging to different test statistics can certainly not be compared in such a way. Instead, one has to compare the corresponding power functions, which is what we shall do next.

The power functions are shown in Figure 1 (“strongly unbalanced,” $n_1 = 3$), Figure 2 (“moderately unbalanced,” $n_1 = 9$), and Figure F.1 (“balanced,” $n_1 = 15$), where only the first two figures are shown in the main text, and the last figure (in which the power functions of all the feasible tests lie “on top of each other”) is available in Appendix F.3 of the Supplementary Material. Readers are referred to the online version for colored figures.

The power functions illustrate that the testing problem is getting easier (i.e., power gets closer to the oracle benchmark), for more balanced design, which has intuitive appeal. Except for the strongly unbalanced case ($n_1 = 3$), the power loss of the tests based on HC0–HC4 and HC0R–HC4R relative to the oracle benchmark is surprisingly small (see Figure 2 as well as Figure F.1 in Appendix F.3 of the Supplementary Material). In the unbalanced cases ($n_1 \in \{3, 9\}$), the HC0–HC4-

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

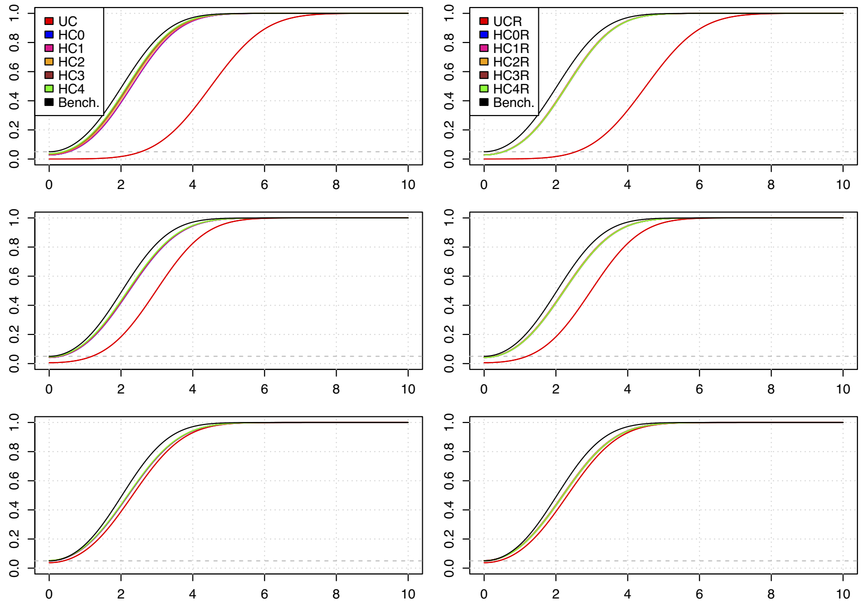

FIGURE 2. Power functions for $n_1 = 9$. Left column: tests based on unrestricted residuals (cf. legend). Right column: tests based on restricted residuals (cf. legend). The rows correspond to Σ_a for $a = 1, 5, 9$ from top to bottom. The abscissa shows δ. In the left panel, the HC0–HC4 curves turn out to be barely distinguishable, with the HC1 curve lying on top of the HC0 curve. In the right panel, the HC4R curve lies on top of the HC0R–HC3R curves. See the text for an explanation.

based tests behave all very similarly, with the power functions of the HC0- and HC1-based test being virtually indistinguishable (as they should in view of the before discussed equivalence). The UC-based test shows markedly worse power performance. Similarly, the HC0R–HC4R-based tests have virtually indistinguishable power functions (as they should because of the before discussed equivalence). The UCR-based test again is inferior (and its power function coincides with the one of UC as mentioned before). There appears also to be little difference between basing the test statistics on unrestricted or restricted residuals in this example. In the balanced case, we know that *all* the feasible tests have exactly the same power function in view of our earlier discussion. This is visible in Figure F.1 in Appendix F.3 of the Supplementary Material. Also, the different forms of heteroskedasticity considered seem not to have much effect on the power functions (when expressed as a function of δ), except for UC and UCR in the unbalanced cases.

Hence, within the scenario considered in this section, perhaps the most important conclusion concerning the choice of a test statistic appears to be to avoid UC and UCR. Everything apart from that, i.e., whether one uses unrestricted or restricted residuals to construct the test or which specific heteroskedasticity

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

correction one decides to use, seems to be a comparably irrelevant part of the problem once the right (i.e., smallest size-controlling) critical value is used. We shall see in the next subsection that this conclusion very much depends on the scenario considered here and does not generalize beyond, illustrating the danger of drawing conclusions from a limited numerical study.

11.2.2. *A High-Leverage Design Matrix.* In this section, we consider testing $\beta_2 = 0$ in a model with intercept and a single regressor $x = (10, \cos(2), \cos(3), \dots, \cos(n))'$. Obviously, the regressor has a dominant first coordinate, leading to diagonal elements h_{ii} of $X(X'X)^{-1}X'$ such that the ratio of largest to smallest h_{ii} is roughly 26 ($\max h_{ii} \simeq 0.879$, $\min h_{ii} \simeq 0.033$). Hence, the design matrix X provides (on purpose) an extreme case, which leads to quite interesting results. We consider again the case $n = 30$ and $\alpha = 0.05$, but now show power functions for Σ_a^*, $a = 0, \dots, 4$, where

$$\Sigma_a^* = n^{-1}\operatorname{diag}\left(7a+1, \frac{n-7a-1}{n-1}, \dots, \frac{n-7a-1}{n-1}\right) \in \mathfrak{C}_{Het}.$$

Note that $\Sigma_0^* = n^{-1}I_n$ and that increasing a from 0 to 4 leads to covariance matrices that approach the degenerate matrix $e_1(n)e_1(n)'$. All conditions in Theorems 5.1 and 6.4 are seen to be satisfied in this example: as no vector $e_i(n)$ belongs to $\operatorname{span}(X)$ (and thus also not to $\mathfrak{M}_0^{lin}$), Assumptions 1 and 2 as well as the sufficient condition for size control (8) are obviously satisfied. The size control conditions (10) and (17) have been checked numerically, as has been the condition that none of the test statistics HC0R–HC4R is constant on $\mathbb{R}^n \backslash \tilde{\mathsf{B}}$.

As in the preceding subsection, the critical values for each test statistic are again chosen as *the smallest critical value guaranteeing size control* over $\mathfrak{C}_{Het}$ and they are presented in Table 5. (Existence follows from our theory since all assumptions are satisfied as noted before.) For their computation, the same algorithms were used as in Section 11.2.1, with a similar statement applying to the numerical routines used for computing the power functions. Note that the critical values for the test statistics UC and HC0–HC3 are large, reflecting the high leverage in the design matrix; an exception is HC4, the reason being that some of the HC4 weights are considerably larger than the weights for HC0–HC3. Similarly as in the preceding subsection, the tests based on HC0 and HC1 coincide (since HC0 and HC1 differ only by a multiplicative constant and since smallest size-controlling critical values are being used), and the same is true for the tests based on HC0R–HC4R (see Remarks 5.3 and 6.5). It is easily checked that the ratios of the respective critical values provided in Table 5 are in good agreement with the theoretical ratios predicted by theory. Furthermore, the tests based on UC and UCR coincide (see Section 6.2.1), and the critical values for UC and UCR in Table 5 are in excellent agreement with their theoretical relationship found in Section 6.2.1.

Table 5 shows that, in this example, the smallest size-controlling critical values are—except in one case—always larger, sometimes considerably larger, than $C_{\chi^2,0.05} \approx 3.8415$, once more showing that the latter critical value is not effecting

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

TABLE 5. Smallest size-controlling critical values for the high-leverage design matrix.

UC	HC0	HC1	HC2	HC3	HC4
217.58	355.56	333.31	121.89	29.34	1.12
UCR	HC0R	HC1R	HC2R	HC3R	HC4R
25.69	5.41	5.45	5.34	5.29	5.44

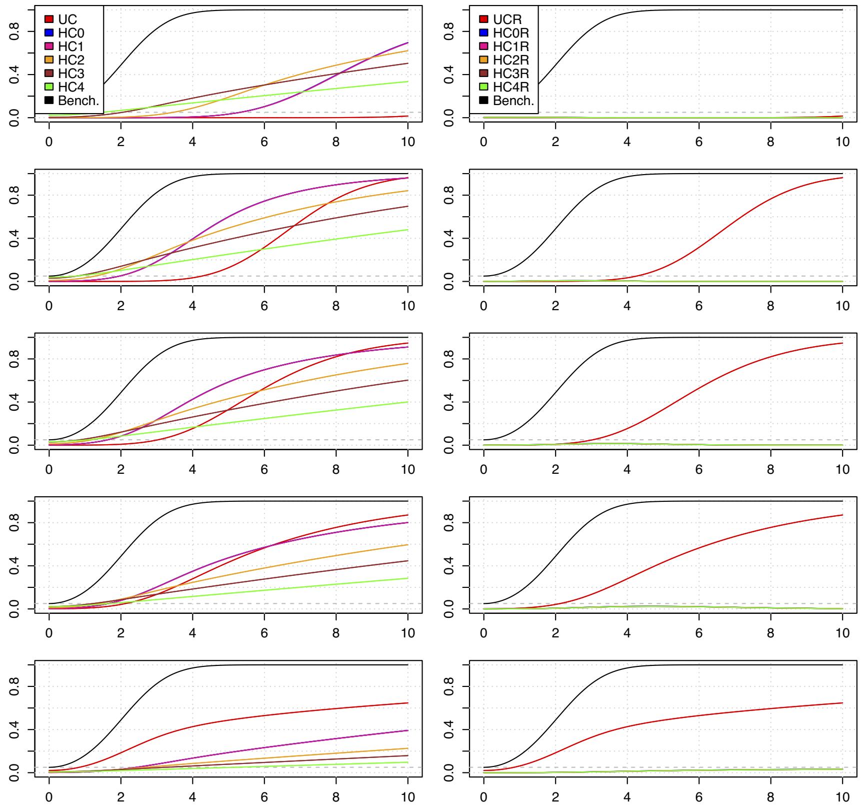

FIGURE 3. Power functions for the design matrix considered in Section 11.2.2. Left column: tests based on unrestricted residuals (cf. legend). Right column: tests based on restricted residuals (cf. legend). The rows from top to bottom correspond to Σ_a^* for $a = 0, 1, 2, 3, 4$, the case $a = 0$ corresponding to homoskedasticity. The abscissa shows δ. In the left panel, the HC1 curve lies on top of the HC0 curve. In the right panel, the HC4R curve lies on top of the HC0R–HC3R curves. See the text for an explanation.

size control in general. In the exceptional case, namely when the HC4 test statistic is used, $C_{\chi^2,0.05}$ is considerably larger than the smallest size-controlling critical value, which is 1.12; while in this case $C_{\chi^2,0.05}$ effects size control, using the

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

smaller size-controlling critical value 1.12 can only be advantageous in terms of power.

The power functions, when the size-controlling critical values from Table 5 are being used, are shown in Figure 3. Readers are referred to the online version for a colored figure. Again, as predicted by theory, the power functions of the tests based on HC0 and HC1 shown in Figure 3 coincide, as do the power functions of the tests based on HC0R–HC4R; the same is true for the power functions of the tests based on UC and UCR. The figure furthermore shows that in the setting considered here, there is now a marked difference between tests based on HC0–HC4 and on HC0R–HC4R, respectively: the power of the tests based on HC0R–HC4R is nowhere greater than α, their power function being even non-monotonic, whereas the tests based on HC0–HC4 have increasing power as a function of δ. In contrast to the example considered in the preceding subsection, the power functions of the tests based on HC0–HC4 and UC are now all markedly different and typically intersect, an exception being the case of Σ_4^* where the test based on UC offers the highest power for that covariance matrix. Overall, however, there is no clear ranking between the tests using unrestricted residuals in the example considered here, although we note that the test based on UC (or, equivalently, on UCR) performs very badly in the case of Σ_0^*. This is not surprising as Σ_0^* corresponds to homoskedasticity and the critical value used here is much larger than the classical critical value one would use given knowledge of this homoskedasticity. Furthermore, and in contrast to the results in the preceding subsection, the different forms of heteroskedasticity considered have a noticeable effect on the power functions. The main takeaway is that tests based on HC0R–HC4R (and probably on UC and UCR) should rather be avoided.

12. CONCLUSION

The usual heteroskedasticity robust test statistics such as T_{Het} (using HC0–HC4 weights) or $\tilde{T}_{Het}$ (using HC0R–HC4R weights), used in conjunction with conventional critical values obtained from the asymptotic null distribution, are often plagued by overrejection under the null. This has been clearly documented in the literature for T_{Het}, and is shown numerically for $\tilde{T}_{Het}$ (as well as for T_{Het}) in Section 11. Not surprisingly, similar observations apply to the "uncorrected" test statistics T_{uc} and $\tilde{T}_{uc}$. We show theoretically that all these test statistics can be size-controlled under quite weak conditions by an appropriate choice of critical values.

From the above discussion and the numerical results in Section 11, it transpires that smallest size-controlling critical values rather than conventional critical values should be used in order to avoid the risk of overrejection. For the computation of smallest size-controlling critical values, we provide algorithms which have been implemented in the R package **hrt** (Preinerstorfer, 2021) and thus are readily available for the user.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

An additional advantage from using smallest size-controlling critical values over conventional critical values is that this typically leads to improved power in instances where conventional critical values lead to underrejection (i.e., lead to worst-case rejection probability under the null less than the nominal significance level) as is sometimes the case (see Sections 6.2.2 and 11.2).

If smallest size-controlling critical values are adopted (as they should), the numerical results in Section 11 suggest that the test statistic $\tilde{T}_{Het}$ (with the usual weights HC0R–HC4R) should be avoided, as the resulting tests may have very poor power properties (see the example in Section 11.2.2). The test statistic T_{Het} seems to perform better in terms of power, with no clear ranking emerging with regard to the weights HC0–HC4 being used. The "uncorrected" test statistics T_{uc} and $\tilde{T}_{uc}$ appear to be inferior to T_{Het} in terms of power in almost all of the numerical examples considered. We also point out that—when using smallest size-controlling critical values—the tests based on T_{Het} employing the HC0 and HC1 weights, respectively, in fact coincide, and the same holds for tests based on $\tilde{T}_{Het}$ employing the HC0R and HC1R weights, respectively. Also, the tests based on T_{uc} and $\tilde{T}_{uc}$ then (essentially) coincide. See Remarks 5.3 and 6.5 and Section 6.2.1 as well as the pertaining discussion in Section 11 for more information, including additional equivalencies when the design matrix X and the restriction R have certain special properties.

SUPPLEMENTARY MATERIAL

Pötscher, B.M. and Preinerstorfer, D. (2023): Supplement to "Valid Heteroskedasticity Robust Testing," Econometric Theory Supplementary Material. To view, please visit: https://doi.org/10.1017/S0266466623000269

References

Bakirov, N. & G. Székely (2005) Student's t-test for Gaussian scale mixtures. *Zapiski Nauchnyh Seminarov POMI* 328, 5–19.

Bakirov, N.K. (1998) Nonhomogeneous samples in the Behrens–Fisher problem. *Journal of Mathematical Sciences (New York)* 89, 1460–1467.

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

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

Chesher, A. & I. Jewitt (1987) The bias of a heteroskedasticity consistent covariance matrix estimator. *Econometrica* 55, 1217–1222.

Chesher, A.D. (1989) Hájek inequalities, measures of leverage, and the size of heteroskedasticity robust Wald tests. *Econometrica* 57, 971–977.

Chesher, A.D. & G. Austin (1991) The finite-sample distributions of heteroskedasticity robust Wald statistics. *Journal of Econometrics* 47, 153–173.

Chu, J., T.-H. Lee, A. Ullah, & H. Xu (2021) Exact distribution of the F-statistic under heteroskedasticity of unknown form for improved inference. *Journal of Statistical Computation and Simulation* 91, 1782–1801.

Cragg, J.G. (1983) More efficient estimation in the presence of heteroscedasticity of unknown form. *Econometrica* 51, 751–763.

Cragg, J.G. (1992) Quasi-Aitken estimation for heteroscedasticity of unknown form. *Journal of Econometrics* 54, 179–201.

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

Davidson, R. & E. Flachaire (2008) The wild bootstrap, tamed at last. *Journal of Econometrics* 146, 162–169.

Davidson, R. & J.G. MacKinnon (1985) Heteroskedasticity-robust tests in regressions directions. *Ministère de l'Économie et des Finances. Institut National de la Statistique et des Études Économiques. Annales* 59/60, 183–218.

Davies, R.B. (1980) Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 29, 323–333.

DiCiccio, C.J., J.P. Romano, & M. Wolf (2019) Improving weighted least squares inference. *Econometrics and Statistics* 10, 96–119.

Duchesne, P. & P.L. de Micheaux (2010) Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Computational Statistics and Data Analysis* 54, 858–862.

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

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Berkeley, CA, 1965/66), Volume 1: Statistics*, pp. 59–82. University of California Press.

Flachaire, E. (2005) More efficient tests robust to heteroskedasticity of unknown form. *Econometric Reviews* 24, 219–241.

Godfrey, L.G. (2006) Tests for regression models with heteroskedasticity of unknown form. *Computational Statistics & Data Analysis* 50, 2715–2733.

Hansen, B. (2021). The exact distribution of the White *t*-ratio. Working paper, University of Wisconsin–Madison.

Hinkley, D.V. (1977) Jackknifing in unbalanced situations. *Technometrics* 19, 285–292.

Ibragimov, R. & U.K. Müller (2010) *t*-statistic based correlation and heterogeneity robust inference. *Journal of Business and Economic Statistics* 28, 453–468.

Ibragimov, R. & U.K. Müller (2016) Inference with few heterogeneous clusters. *The Review of Economics and Statistics* 98, 83–96.

Imbens, G.W. & M. Kolesár (2016) Robust standard errors in small samples: Some practical advice. *The Review of Economics and Statistics* 98, 701–712.

Kolesár, M. (2019). *dfadjust: Degrees of Freedom Adjustment for Robust Standard Errors*. R package version 1.0.1. https://CRAN.R-project.org/package=dfadjust.

Lin, E.S. & T.-S. Chou (2018) Finite-sample refinement of GMM approach to nonlinear models under heteroskedasticity of unknown form. *Econometric Reviews* 37, 1–28.

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

MacKinnon, J.G. (2013) Thirty years of heteroskedasticity-robust inference. In X. Chen & N. R. E. Swanson (eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pp. 437–462. Springer.

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

Mickey, M.R. & M.B. Brown (1966) Bounds on the distribution functions of the Behrens–Fisher statistic. *Annals of Mathematical Statistics* 37, 639–642.

Phillips, P.C. (1993) Operational algebra and regression *t*-tests. In P.C. Phillips (ed.), *Models, Methods and Applications of Econometrics: Essays in Honor of A.R. Bergstrom*, pp. 140–152. Basil Blackwell.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press

Pötscher, B.M. & D. Preinerstorfer (2018) Controlling the size of autocorrelation robust tests. *Journal of Econometrics* 207, 406–431.

Pötscher, B.M. & D. Preinerstorfer (2019) Further results on size and power of heteroskedasticity and autocorrelation robust tests, with an application to trend testing. *Electronic Journal of Statistics* 13, 3893–3942.

Pötscher, B.M. & D. Preinerstorfer (2023) How reliable are bootstrap-based heteroskedasticity robust tests? *Econometric Theory* 39(4), 789–847. doi:10.1017/S0266466622000184.

Preinerstorfer, D. (2021). *hrt: Heteroskedasticity Robust Testing*. R package version 1.0.0.

Preinerstorfer, D. & B.M. Pötscher (2016) On size and power of heteroskedasticity and autocorrelation robust tests. *Econometric Theory* 32, 261–358.

Robinson, G. (1979) Conditional properties of statistical procedures. *Annals of Statistics* 7, 742–755.

Romano, J.P. & M. Wolf (2017) Resurrecting weighted least squares. *Journal of Econometrics* 197, 1–19.

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

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

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

Wooldridge, J.M. (2010) *Econometric Analysis of Cross Section and Panel Data*, *2nd Edition*. MIT Press.

Wooldridge, J.M. (2012) *Introductory Econometrics*, *5th Edition*. South-Western.

https://doi.org/10.1017/S0266466623000269 Published online by Cambridge University Press