Testing Many Restrictions Under Heteroskedasticity

Stanislav Anatolyev[*], Mikkel Sølvsten[†]

December 2020

Abstract

We propose a hypothesis test that allows for many tested restrictions in a heteroskedastic linear regression model. The test compares the conventional F statistic to a critical value that corrects for many restrictions and conditional heteroskedasticity. The correction utilizes leave-one-out estimation to correctly center the critical value and leave-three-out estimation to appropriately scale it. Large sample properties of the test are established in an asymptotic framework where the number of tested restrictions may be fixed or may grow with the sample size and can even be proportional to the number of observations. We show that the test is asymptotically valid and has non-trivial asymptotic power against the same local alternatives as the exact F test when the latter is valid. Simulations corroborate the relevance of these theoretical findings and suggest excellent size control in moderately small samples also under strong heteroskedasticity.

Keywords: linear regression, ordinary least squares, many regressors, leave-out estimation, hypothesis testing, high-dimensional models.

JEL codes: C12, C13, C21

[*]CERGE-EI, Politických vězňů 7, 11121 Prague 1, Czech Republic. E-mail: stanislav.anatolyev@cerge-ei.cz. This research was supported by the grant 20-28055S from the Czech Science Foundation.

[†]Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706, United States. E-mail: soelvsten@wisc.edu.

1 Introduction

One of the central tenets in modern economic research is to consider models that allow for flexible specifications of heterogeneity and to establish whether there is the presence or absence of meaningful heterogeneity in particular empirical settings. For example, Abowd et al. (1999) studies whether there is firm-specific heterogeneity in a linear model for individual log-wages, Card et al. (2018) asks if this heterogeneity varies by the individual's education, and Lachowska et al. (2019) investigates whether the firm-specific heterogeneity is constant over time. Other work relies on similarly flexible models to investigate the presence of heterogeneity in health economics (Finkelstein et al., 2016), in educational settings (Sacerdote, 2001), and to study neighborhood effects (Chetty and Hendren, 2018). In all these examples, the absence of a particular dimension of heterogeneity corresponds to a hypothesis that imposes hundreds or thousands of restrictions on the model of interest. The present paper provides a tool to conduct a test of such hypotheses.

We develop a test for hypotheses that impose multiple restrictions, and establishe its asymptotic validity in a heteroskedastic linear regression model where the number of tested restrictions may be fixed or increasing with the sample size. In particular, we allow for the number of restrictions and the sample size to be proportional. The exact F test fails to control size in this environment, so our proposed test instead rejects the hypothesis if the F statistic exceeds a critical value that corrects for many restrictions and conditional heteroskedasticity. This critical value is a recentered and rescaled quantile of what is naturally called the *F-bar distribution* as it describes the distribution of a chi-bar-squared random variable divided by an independent chi-squared random variable over its degrees of freedom. This family of distributions can approximate both the finite sample properties of the F statistic in the special case of homoskedastic normal errors and—after recentering and rescaling—the asymptotic distribution of the F statistic in the presence of conditional heteroskedasticity and few or many restrictions.

The location and variance estimators used to recenter and rescale the critical value utilize unbiased leave-one-out estimators for individual error variances and unbiased leave-*three*-out

estimators for *products* of individual error variances. While leave-one-out estimation is used in many other econometric contexts, introduction of leave-three-out estimation is entirely new in the literature. Because these essential elements of the test are built on *l*eave-*o*ut machinery, we will at times and for brevity refer to the proposed test using the acronym LO.

The LO test has exact asymptotic size when the regression design has full rank after leaving any combination of three observations out of the sample. This condition is, in general, satisfied in models with many continuous regressors and only a few discrete ones. However, the condition may fail when many discretely valued regressors are included, as occurs for models with one or more fixed (or group) effects. To handle such cases, our proposed test uses estimators for the products of individual error variances that are intentionally biased upward when the unbiased leave-three-out estimator fails to exist. This construction ensures validity in large samples but can potentially be slightly conservative when many of the leave-three-out estimators do not exist.

Using a combination of theoretical arguments and simulations, Huber (1973) and Berndt and Savin (1977) highlight the importance of allowing the number of regressors and potentially the number of tested restrictions to increase with sample size when studying asymptotic properties of inference procedures. The latter specifically documents conflicts among classical tests when the number of tested restrictions is somewhat large. Despite these early cautionary tales, most inference procedures that allow for proportionality between the number of regressors, sample size, and potentially the number of restrictions, are of a more recent vintage. Here, we survey the ones most relevant to the current paper and refer to Anatolyev (2019) for a more extensive review of the literature.[1]

In homoskedastic regression models, Anatolyev (2012) and Calhoun (2011) propose various corrections to classical tests that restore asymptotic validity in the presence of many restrictions. In heteroskedastic regression models with one tested restriction and many regressors, Cattaneo et al. (2018) show that the use of conventional Eicker-White standard

[1] In analysis of variance contexts, which are special cases of linear regression, Akritas and Papadatos (2004) and Zhou et al. (2017) propose heteroskedasticity robust tests for equality of means that are, however, specific to their models. An expanding literature considers (outlier) robust estimation of linear high-dimensional regressions (e.g., El Karoui et al., 2013) but does not provide valid tests of many restrictions.

errors and their "almost-unbiased" variations (see MacKinnon, 2013) does not yield asymptotic validity. This failure may be viewed as a manifestation of the incidental parameters problem. To overcome this problem, Cattaneo et al. (2018) and subsequently Anatolyev (2018) propose new versions of the Eicker-White standard errors, which restore size control in large samples. However, these proposals rely on the inversion of n-by-n matrices (n denotes sample size) that may fail to be invertible in examples of practical interest (Horn et al., 1975; Verdier, 2020). Rao (1970)'s unbiased estimator for individual error variances is closely related to Cattaneo et al. (2018)'s proposal and suffers from the same existence issue.

Kline, Saggio, and Sølvsten (2020) propose instead a version of the Eicker-White standard errors that relies only on leave-one-out estimators of individual error variances and show that its use leads to asymptotic size control when testing a single restriction.[2] While this conclusion extends to hypotheses that involve a fixed and small number of restrictions through the use of a heteroskedasticity-robust Wald test, it can fail to hold in cases of many restrictions. When testing many coefficients equal to zero, Kline et al. (2020) note instead that those leave-one-out individual variance estimators can be used to center the conventional F statistic[3] and propose a rescaling of the statistic that relies on successive sample splitting. However, outside of the specific model of interest in their empirical application, sample splitting places restrictions on the data that may often fail in practice.

Here we propose a feasible scaling of the critical value that uses a leave-three-out approach, which requires less from the data than sample splitting. Additionally, we propose a one-shot test that enables asymptotic size control both when the number of restrictions is fixed and increasing. Finally, we provide a theoretical study of the power properties under local and global alternatives and conduct a simulation study that documents the performance

[2] Jochmans (2020) additionally uses simulations to investigate the finite sample behavior of this variance estimator.

[3] The use of leave-one-out estimation has a long tradition in the literature on instrumental variables (see, e.g., Phillips and Hale, 1977), and our test shares an algebraic representation with the adjusted J test analyzed in Chao et al. (2014) (see Kline et al., 2020, for a discussion). An attractive feature of relying on leave-one-out is that challenging estimation of higher order error moments can be avoided, which is in contrast to the tests of Calhoun (2011) and Anatolyev (2013).

of our test in small and moderately sized samples.

Under local alternatives, the asymptotic power curve of the proposed LO test is parallel to that of the exact F test when the latter is valid, e.g., under homoskedastic normal errors. While the curves are parallel, the LO test tends to have power somewhat below the exact F test. This loss in power stems from the estimation of individual error variances and can be viewed as a cost of using a test that is robust to general heteroskedasticity. This cost is largely monotone in the number of tested restrictions and disappears when the number of restrictions is small relative to sample size.

Using a simulation study, we document that the LO test delivers a nearly exact size control in samples as small as 100 observations in both homoskedastic and heteroskedastic environments. On the other hand, conventional tools such as the Wald test and the exact F test can exhibit severe size distortions and reject a true null with near certainty for some configurations. These findings are documented using two simulation settings: one with continuous regressors only, and one with a mix of both continuous and discrete regressors. In the latter setting, roughly 7% of observations cause a full rank failure when leaving up to three observations out, but the proposed test shows almost no conservatism even in this adverse environment. When both the LO and exact F tests are valid, the simulations document a power loss that varies between being negligible and up to roughly 15 percentage points, depending on the type of deviation from the null and sample size. For many applications, this range of power losses is a small cost to incur for being robust to heteroskedasticity.

The paper is organized as follows. Section 2 introduces the setup and the proposed critical value in samples where all the leave-three-out estimators exist, while Section 3 analyzes the asymptotic size and power of the LO test for such samples. Section 4 describes the critical value for use in samples where the design loses full rank after leaving certain triples of observations out. Section 5 discusses the results of simulation experiments, and Section 6 concludes. Proofs of theoretical results and some clarifying but technical details are collected in the online supplemental Appendix. An R package (Anatolyev and Sølvsten, 2020) that implements the proposed test is available online.

2 Leave-out test

Consider a linear regression model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i|\boldsymbol{x}_i] = 0,$$

where an intercept is included in the regression function $\boldsymbol{x}_i'\boldsymbol{\beta}$ and the n observed random vectors $\{(y_i, \boldsymbol{x}_i')'\}_{i=1}^n$ are independent across i. The dimension of the regressors $\boldsymbol{x}_i \in \mathbb{R}^m$ may be large relative to sample size, and there is conditional heteroskedasticity in the unobserved errors: $\mathbb{E}[\varepsilon_i^2|\boldsymbol{x}_i] = \sigma^2(\boldsymbol{x}_i) \equiv \sigma_i^2$. The hypothesis of interest involves $r \leq m$ linear restrictions

$$H_0 \ : \ \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{q},$$

where the matrix $\boldsymbol{R} \in \mathbb{R}^{r\times m}$ has full row rank r, and $\boldsymbol{q} \in \mathbb{R}^r$. Both $\boldsymbol{R}$ and $\boldsymbol{q}$ are specified by the researcher. Specifically, they are assumed to be known and are allowed to depend on the observed regressors. The space of alternatives is $H_A \ : \ \boldsymbol{R}\boldsymbol{\beta} \neq \boldsymbol{q}$.

The attention of the paper is on settings where the design matrix $\boldsymbol{S}_{xx} = \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'$ has full rank so that $\hat{\boldsymbol{\beta}} = \boldsymbol{S}_{xx}^{-1}\sum_{i=1}^n \boldsymbol{x}_i y_i$, the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$, is defined. For compact reference, we define the degrees-of-freedom adjusted residual variance $\hat{\sigma}_\varepsilon^2 = (n-m)^{-1}\sum_{i=1}^n \left(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\right)^2$. Unless otherwise noted, all means and variances are conditional on the regressors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, and the means and variances with a subscript 0 are computed under H_0.

2.1 Test statistic

Our proposed test rejects H_0 for large values of Fisher's F statistic,

$$F = \frac{\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{q}\right)'\left(\boldsymbol{R}\boldsymbol{S}_{xx}^{-1}\boldsymbol{R}'\right)^{-1}\left(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{q}\right)}{r\hat{\sigma}_\varepsilon^2},$$

which is a monotone transformation of the likelihood ratio statistic when the regression errors are homoskedastic normal. Since we do not impose normality, F may be viewed as a quasi likelihood ratio statistic.

By taking this statistic as a point of departure, we are able to construct a critical value that ensures size control in the presence of heteroskedasticity and any number of restrictions. An alternative approach might have taken a heteroskedasticity-robust Wald statistic and attempted to construct a critical value that ensures validity even when r is proportional to n. However, in such environments, any heteroskedasticity-robust Wald statistic relies on the inverse of a high-dimensional covariance matrix estimator, a feature that presents substantial challenges when attempting to control size. For this reason, we leave the theoretical investigation of such Wald statistics to future research, but note that typically used tests based on Wald statistics exhibit extreme size distortions in our simulation experiments.

Our proposed critical value for the F statistic achieves asymptotic validity under two asymptotic frameworks, one where the number of restrictions is fixed, and one where the number of restrictions may grow as fast as proportionally to the sample size. To achieve such uniformity with respect to the number of restrictions, we rely on an auxiliary distribution, the F-bar distribution, that helps unite these two frameworks.

2.2 F-bar distribution

Our test rejects H_0 if Fisher's F exceeds a linearly transformed quantile of a distribution, which we call the *F-bar distribution*. We define this family of distributions and discuss its role before we turn to a description of the linear transformation mentioned above.

Definition 1 (F-bar distribution)**.** Let $\boldsymbol{w} = (w_1, \ldots, w_r)$ be a collection of non-negative weights summing to one, and df be a positive real number. The F-bar distribution with weights $\boldsymbol{w}$ and degrees of freedom df, denoted by $\bar{F}_{\boldsymbol{w},df}$, is a distribution of

$$\frac{\sum_{\ell=1}^{r} w_\ell Z_\ell}{Z_0/df}, \tag{1}$$

where $Z_0, Z_1, \ldots, Z_r$ are mutually independent random variables with $Z_0 \sim \chi^2_{df}$ and $Z_\ell \sim \chi^2_1$ for $1 \leq \ell \leq r$. Here, χ^2_κ denotes a chi-squared distribution with $\kappa > 0$ degrees of freedom.

The name attached to this family originates from its close relationship to both the chi-bar-squared distribution and to Snedecor's F distribution, which we denote as $\bar{\chi}^2_{\boldsymbol{w}}$ and $F_{r,df}$, respectively. Snedecor's F is a special case when the entries of $\boldsymbol{w}$ are all equal, while $\bar{\chi}^2_{\boldsymbol{w}}$ is a limiting case when $df \to \infty$. Another essential property of this family is that the standard normal distribution, denoted Φ, is also a limiting case since, as $df \to \infty$ and $\max_{1\leq\ell\leq r} w_\ell \to 0$,

$$\frac{q_\tau(\bar{F}_{\boldsymbol{w},df}) - 1}{\sqrt{2\sum_{\ell=1}^r w_\ell^2 + 2/df}} \to q_\tau(\Phi) \tag{2}$$

for $\tau \in (0,1)$, where $q_\tau(G)$ denotes the τ-th quantile of the distribution G. The centering and rescaling in (2) are done according to the limiting mean and variance of the underlying random variable from Definition 1 following the $\bar{F}_{\boldsymbol{w},df}$ distribution, while asymptotic normality results from mixing over infinitely many independent chi-squared variables.

Our reliance on the F-bar distribution is tied to its three properties described in the previous paragraph and three closely related observations about the F statistic. These observations are: (i) the F statistic is distributed as $F_{r,n-m}$ if the errors are homoskedastic normal, (ii) the F statistic converges in distribution (after rescaling) to a χ-bar-squared if the number of restrictions r is fixed, and (iii) the F statistic converges in distribution (after centering and rescaling) to a standard normal as r grows. Therefore, the class of F-bar distributions serves as a roof designed both to match the finite sample distribution of the F statistic in an important special case and to approximate each of the possible limiting distributions after a suitable linear transformation.

2.3 Critical value

The proposed critical value at a nominal size $\alpha \in (0,1)$ is a linear transformation of $q_{1-\alpha}\big(\bar{F}_{\hat{\boldsymbol{w}},n-m}\big)$ given by

$$c_\alpha = \frac{1}{r\hat{\sigma}_\varepsilon^2}\left(\hat{E}_\mathcal{F} + \hat{V}_\mathcal{F}^{1/2}\frac{q_{1-\alpha}(\bar{F}_{\hat{\boldsymbol{w}},n-m}) - 1}{\sqrt{2\sum_{\ell=1}^r \hat{w}_\ell^2 + 2/(n-m)}}\right).$$

The data-dependent quantities $\hat{E}_\mathcal{F}$ and $\hat{V}_\mathcal{F}$ are related to the numerator of the F statistic, which we denote by $\mathcal{F}$. In particular, $\hat{E}_\mathcal{F}$ is an unbiased estimator of the conditional mean $\mathbb{E}_0[\mathcal{F}]$, while $\hat{V}_\mathcal{F}$ is either an unbiased or positively biased estimator of the conditional variance $\mathbb{V}_0[\mathcal{F} - \hat{E}_\mathcal{F}]$ as explained further below. The estimated weights $\hat{\boldsymbol{w}} = (\hat{w}_1, \ldots, \hat{w}_r)$ are constructed to be consistent for weights $\boldsymbol{w}_\mathcal{F}$ in those cases where $\mathcal{F}/\mathbb{E}_0[\mathcal{F}]$ converges in distribution to $\bar{\chi}^2_{\boldsymbol{w}_\mathcal{F}}$.

The critical value c_α ensures asymptotic size control irrespective of whether r is viewed as fixed or growing with the sample size n. To explain why c_α provides such uniformity, we consider first the case where r grows. In this case, it is illuminating to rewrite the rejection rule as an equivalent event

$$\hat{V}_\mathcal{F}^{-1/2}(\mathcal{F} - \hat{E}_\mathcal{F}) > \frac{q_{1-\alpha}(\bar{F}_{\hat{\boldsymbol{w}},n-m}) - 1}{\sqrt{2\sum_{\ell=1}^r \hat{w}_\ell^2 + 2/(n-m)}}.$$

Since $\hat{V}_\mathcal{F}^{-1/2}(\mathcal{F} - \hat{E}_\mathcal{F})$ is asymptotically normal under the null, the validity in large samples follows from the relationship between the F-bar and standard normal distributions given in (2).

When instead r is viewed as asymptotically fixed, it is more informative to express the rejection region through the inequality

$$\frac{\mathcal{F}}{\hat{E}_\mathcal{F}} > q_{1-\alpha}(\bar{F}_{\hat{\boldsymbol{w}},n-m}) + \big(q_{1-\alpha}(\bar{F}_{\hat{\boldsymbol{w}},n-m}) - 1\big)\left(\frac{\hat{V}_\mathcal{F}^{1/2}/\hat{E}_\mathcal{F}}{\sqrt{2\sum_{\ell=1}^r \hat{w}_\ell^2 + 2/(n-m)}} - 1\right). \tag{3}$$

Note that rejecting when $\mathcal{F}/\hat{E}_{\mathcal{F}}$ exceeds the quantile $q_{1-\alpha}(\bar{F}_{\hat{\boldsymbol{w}},n-m})$ suffices for validity; for the case of a single restriction such an approach corresponds to the standard practice of comparing squares of a heteroskedasticity robust t statistic and the $(1-\alpha)$-th quantile of Student's t distribution with $n-m$ degrees of freedom.[4] The last term on the right hand side of (3) can then be viewed as a finite sample correction that adjusts the critical value up or down depending on the relative size of the variance estimator for the ratio $\mathcal{F}/\hat{E}_{\mathcal{F}}$, which is $\hat{V}_{\mathcal{F}}/\hat{E}_{\mathcal{F}}^2$, and the variance of the approximating distribution $\bar{F}_{\hat{\boldsymbol{w}},n-m}$, which is roughly $2\sum_{\ell=1}^{r}\hat{w}_\ell^2 + 2/(n-m)$. As the ratio of these variances converges to unity when the number of restrictions is fixed, this term does not affect first order asymptotic validity.

Finally, note that if one is willing to rest on the assumption that the restrictions are numerous and the few restriction framework is superfluous, one might use the following simplified critical value not robust to few restrictions:[5]

$$\check{c}_\alpha = \frac{1}{r\hat{\sigma}_\varepsilon^2}\left(\hat{E}_{\mathcal{F}} + \hat{V}_{\mathcal{F}}^{1/2}\frac{q_{1-\alpha}(F_{r,n-m})-1}{\sqrt{2/r+2/(n-m)}}\right).$$

To complete the description of the proposed critical value, definitions of the quantities $\hat{E}_{\mathcal{F}}$, $\hat{V}_{\mathcal{F}}$ and $\hat{\boldsymbol{w}}$ are needed. Section 2.5 describes how we rely on leave-one-out OLS estimators to construct $\hat{E}_{\mathcal{F}}$ and $\hat{\boldsymbol{w}}$. For $\hat{V}_{\mathcal{F}}$, Section 2.6 provides the corresponding definition when it is possible to rely on leave-three-out OLS estimators, while Section 4 introduces the form of $\hat{V}_{\mathcal{F}}$ for settings where some of the leave-three-out estimators cease to exist. In the former case, it is possible to ensure that $\hat{V}_{\mathcal{F}}$ is unbiased, while the latter introduces a (small) positive bias. We initially consider the former case, a framework where the design matrix has full rank when any three observations are left out of the sample, and relax this condition in Section 4.

Assumption 1. $\sum_{\ell\neq i,j,k}\boldsymbol{x}_\ell\boldsymbol{x}_\ell'$ *is invertible for every* $i,j,k\in\{1,\ldots,n\}$.

When $\boldsymbol{x}_i$ is identically and continuously distributed with unconditional second moment

[4] When testing a single restriction, $\hat{\boldsymbol{w}}$ must equal unity so that $\bar{F}_{\hat{\boldsymbol{w}},n-m} = F_{1,n-m} = t_{n-m}^2$, and in this case $\mathcal{F}/\hat{E}_{\mathcal{F}}$ is the square of the t statistic studied in Kline et al. (2020, Theorem 1).

[5] Such settings occur, for example, if the null of interest involves thousands of restrictions, in which case the two critical values c_α and $\check{c}_\alpha$ are essentially equivalent but $\check{c}_\alpha$ is computationally simpler to construct as it circumvents computation of $\hat{\boldsymbol{w}}$.

$\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i']$ of full rank, Assumption 1 holds with probability one whenever $n-m \geq 3$. The asymptotic framework considers a setting where $n-m$ diverges so that Assumption 1 must hold in sufficiently large samples with continuous regressors. This conclusion also applies when $\boldsymbol{x}_i$ includes a few discrete regressors and, in particular, an intercept. In settings with many discrete regressors, Assumption 1 may fail to hold, even in large samples. For that reason, Section 4 introduces the version of $\hat{V}_{\mathcal{F}}$ for empirical settings where the full rank condition is satisfied when any one observation is left out, but not necessarily when leaving two or three observations out.

2.4 Leave-out algebra

Before describing $\hat{E}_{\mathcal{F}}$, $\hat{V}_{\mathcal{F}}$, and $\hat{\boldsymbol{w}}$ in detail, we will reformulate Assumption 1 using leave-out algebra. That is, we will derive an equivalent way of expressing this assumption while introducing notation that is essential for the construction of the critical value and for stating the asymptotic regularity conditions.

When $\boldsymbol{S}_{xx}$ has full rank, a direct implication of the Sherman-Morrison-Woodbury identity (Sherman and Morrison, 1950; Woodbury, 1949, SMW) is that the leave-one-out design matrix $\sum_{j\neq i}\boldsymbol{x}_j\boldsymbol{x}_j'$ is invertible if and only if the statistical leverage of the i-th observation $P_{ii} = \boldsymbol{x}_i'\boldsymbol{S}_{xx}^{-1}\boldsymbol{x}_i$ is less than one. Letting $M_{ij} = \mathbf{1}\{i=j\} - \boldsymbol{x}_i'\boldsymbol{S}_{xx}^{-1}\boldsymbol{x}_j$ be elements of the residual projection matrix $\boldsymbol{M}$ associated with the regressor matrix, this condition on the leverage is equivalently stated as M_{ii} being greater than zero. When $M_{ii} > 0$ holds, we can additionally use SMW to represent the inverse of $\sum_{j\neq i}\boldsymbol{x}_j\boldsymbol{x}_j'$ as

$$\left(\sum\nolimits_{j\neq i}\boldsymbol{x}_j\boldsymbol{x}_j'\right)^{-1} = \boldsymbol{S}_{xx}^{-1} + \frac{\boldsymbol{S}_{xx}^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\boldsymbol{S}_{xx}^{-1}}{M_{ii}}, \tag{4}$$

which highlights the role of a non-zero M_{ii}.

The representation in (4) can also be used to understand when the leave-two-out design matrix $\sum_{k\neq i,j}\boldsymbol{x}_k\boldsymbol{x}_k'$ has full rank, since (4) can be used to compute leverages in a sample that excludes i. After leaving observation i out, the leverage of a different observation j is

$\boldsymbol{x}_j'\big(\sum_{k\neq i}\boldsymbol{x}_k\boldsymbol{x}_k'\big)^{-1}\boldsymbol{x}_j$. To see when this leverage is less than one, note that (4) yields

$$1-\boldsymbol{x}_j'\Big(\sum\nolimits_{k\neq i}\boldsymbol{x}_k\boldsymbol{x}_k'\Big)^{-1}\boldsymbol{x}_j=M_{jj}-\frac{M_{ij}^2}{M_{ii}},$$

so that a necessary and sufficient condition for a full rank of $\sum_{k\neq i,j}\boldsymbol{x}_k\boldsymbol{x}_k'$ is that $D_{ij}>0$, where

$$D_{ij}=\begin{vmatrix}M_{ii} & M_{ij}\\ M_{ij} & M_{jj}\end{vmatrix}=M_{ii}M_{jj}-M_{ij}^2,$$

and $|\cdot|$ denotes the determinant. Extending the previous argument to the case of leaving three observations out, we find that the invertibility of $\sum_{\ell\neq i,j,k}\boldsymbol{x}_\ell\boldsymbol{x}_\ell'$ for i, j, and k, all of which are different, is equivalent to $D_{ijk}>0$, where

$$D_{ijk}=\begin{vmatrix}M_{ii} & M_{ij} & M_{ik}\\ M_{ij} & M_{jj} & M_{jk}\\ M_{ik} & M_{jk} & M_{kk}\end{vmatrix}=M_{ii}D_{jk}-\left(M_{jj}M_{ik}^2+M_{kk}M_{ij}^2-2M_{jk}M_{ij}M_{ik}\right).$$

This discussion reveals that Assumption 1 can equivalently be stated as requiring full rank of $\boldsymbol{S}_{xx}$ and

$$D_{ijk}>0 \text{ for every } i,j,k\in\{1,\ldots,n\} \text{ with } i\neq j\neq k\neq i. \tag{5}$$

In addition to facilitating an algebraic description of Assumption 1, the quantities M_{ii}, D_{ij}, and D_{ijk} also play a role in the computation of the proposed critical value. Specifically, they can be used to avoid explicitly computing the OLS estimates after leaving one, two, or three observations out. Additionally, since construction of $\hat{E}_{\mathcal{F}}$, $\hat{V}_{\mathcal{F}}$, and $\hat{\boldsymbol{w}}$ relies on dividing by M_{ii}, D_{ij}, and D_{ijk}, the study of the asymptotic size of the proposed testing procedure imposes a slight strengthening of (5), which bounds the smallest D_{ijk} away from zero.

2.5 Location estimator

Recall that the numerator of the F statistic is

$$\mathcal{F} = \big(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{q}\big)'\big(\boldsymbol{R}\boldsymbol{S}_{xx}^{-1}\boldsymbol{R}'\big)^{-1}\big(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{q}\big).$$

A virtue of this statistic is that its expectation is minimized under H_0, so that large values of the statistic can be taken as evidence against the hypothesized value of $\boldsymbol{\beta}$. However, the distribution of $\mathcal{F}$ depends on the unknown error variances $\{\sigma_i^2\}_{i=1}^n$, which complicates the construction of a critical value. Specifically, the conditional mean of $\mathcal{F}$ under H_0 is

$$\mathbb{E}_0[\mathcal{F}] = \sum_{i=1}^{n} B_{ii}\sigma_i^2,$$

where the values $B_{ij} = \boldsymbol{x}_i'\boldsymbol{S}_{xx}^{-1}\boldsymbol{R}'\big(\boldsymbol{R}\boldsymbol{S}_{xx}^{-1}\boldsymbol{R}'\big)^{-1}\boldsymbol{R}\boldsymbol{S}_{xx}^{-1}\boldsymbol{x}_j$ are observed and satisfy $\sum_{i=1}^n B_{ii} = r$. Furthermore, the exact null distribution of $\mathcal{F}/\mathbb{E}_0[\mathcal{F}]$, under the additional condition of normally distributed regression errors, is $\bar{\chi}^2_{\boldsymbol{w}_\mathcal{F}}$, with $\boldsymbol{w}_\mathcal{F}$ containing the eigenvalues of the matrix[6]

$$\begin{aligned}\Omega\big(\sigma_1^2,\ldots,\sigma_n^2\big) &= \frac{1}{\mathbb{E}_0[\mathcal{F}]}\big(\boldsymbol{R}\boldsymbol{S}_{xx}^{-1}\boldsymbol{R}'\big)^{-1}\mathbb{V}\big[\boldsymbol{R}\hat{\boldsymbol{\beta}}\big] \\ &= \frac{1}{\sum_{i=1}^n B_{ii}\sigma_i^2}\big(\boldsymbol{R}\boldsymbol{S}_{xx}^{-1}\boldsymbol{R}'\big)^{-1}\boldsymbol{R}\boldsymbol{S}_{xx}^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\sigma_i^2\right)\boldsymbol{S}_{xx}^{-1}\boldsymbol{R}'.\end{aligned}$$

Both $\mathbb{E}_0[\mathcal{F}]$ and $\boldsymbol{w}_\mathcal{F}$ are thus functions of $\{\sigma_i^2\}_{i=1}^n$, and the relevance of the vector $\boldsymbol{w}_\mathcal{F}$ for asymptotic size control transcends the normality assumption on the errors that we used in order to introduce it.

As shown in Kline et al. (2020), the individual specific error variances can be estimated without bias for any value of $\boldsymbol{\beta}$ using leave-one-out estimators. Let the leave-i-out OLS

[6]The eigenvalues of $\Omega(\sigma_1^2,\ldots,\sigma_n^2)$ are all real and non-negative as they can be expressed as the eigenvalues of the symmetric and positive semidefinite matrix $(\boldsymbol{R}\boldsymbol{S}_{xx}^{-1}\boldsymbol{R}')^{-1/2}\mathbb{V}\big[\boldsymbol{R}\hat{\boldsymbol{\beta}}\big](\boldsymbol{R}\boldsymbol{S}_{xx}^{-1}\boldsymbol{R}')^{-1/2}/\mathbb{E}_0[\mathcal{F}]$. Furthermore, the entries of $\boldsymbol{w}_\mathcal{F}$ sum to one as $\mathbb{E}_0[\mathcal{F}]$ is the trace of $(\boldsymbol{R}\boldsymbol{S}_{xx}^{-1}\boldsymbol{R}')^{-1}\mathbb{V}\big[\boldsymbol{R}\hat{\boldsymbol{\beta}}\big]$.

estimator of $\boldsymbol{\beta}$ be $\hat{\boldsymbol{\beta}}_{-i} = \big(\sum_{j\neq i} \boldsymbol{x}_j \boldsymbol{x}_j'\big)^{-1} \sum_{j\neq i} \boldsymbol{x}_j y_j$, and construct

$$\hat{\sigma}_i^2 = y_i\big(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{-i}\big).$$

With these leave-one-out estimators, we can estimate the null mean of $\mathcal{F}$ using

$$\hat{E}_{\mathcal{F}} = \sum_{i=1}^{n} B_{ii}\hat{\sigma}_i^2,$$

which ensures that the first moment of $\mathcal{F} - \hat{E}_{\mathcal{F}}$ is zero under the null. Since $\hat{\sigma}_i^2$ is unbiased for any value of $\boldsymbol{\beta}$, this centered statistic still has its expectation minimized under H_0, so that large values of the statistic can be taken as evidence against the null. Following the same approach, we can estimate $\boldsymbol{w}_{\mathcal{F}}$ using the sample analog $\check{\boldsymbol{w}} = (\check{w}_1, \ldots, \check{w}_n)'$, where $\check{w}_\ell$ is the ℓ-th eigenvalue of $\Omega\big(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_n^2\big)$. However, $\check{\boldsymbol{w}}$ may not have non-negative entries summing to one, so we ensure that these conditions hold by letting $\hat{\boldsymbol{w}} = (\hat{w}_1, \ldots, \hat{w}_n)'$, where

$$\hat{w}_\ell = \frac{\check{w}_\ell \vee 0}{\sum_{\ell=1}^{r}(\check{w}_\ell \vee 0)}.$$

While our construction of $\hat{E}_{\mathcal{F}}$ implies that the first moment of $\mathcal{F} - \hat{E}_{\mathcal{F}}$ is known when H_0 holds, its second moment still depends heavily on unknown parameters. Under H_0,

$$\mathbb{V}_0\Big[\mathcal{F} - \hat{E}_{\mathcal{F}}\Big] = \sum_{i=1}^{n}\sum_{j\neq i} U_{ij}\sigma_i^2\sigma_j^2 + \sum_{i=1}^{n}\Big(\sum\nolimits_{j\neq i} V_{ij}\boldsymbol{x}_j'\boldsymbol{\beta}\Big)^2\sigma_i^2, \tag{6}$$

where $U_{ij} = 2\big(B_{ij} - M_{ij}\big(B_{ii}/M_{ii} + B_{jj}/M_{jj}\big)/2\big)^2$ and $V_{ij} = M_{ij}\big(B_{ii}/M_{ii} - B_{jj}/M_{jj}\big)$ are known quantities. This representation of the null-variance stems from writing $\mathcal{F} - \hat{E}_{\mathcal{F}}$ as a second order U-statistic with squared kernel weights of $U_{ij}/2$ plus a linear term with weights $\sum_{j\neq i} V_{ij}\boldsymbol{x}_j'\boldsymbol{\beta}$ (see the Appendix for details).

2.6 Variance estimator

This subsection describes the construction of an unbiased estimator of the conditional variance $\mathbb{V}_0\big[\mathcal{F} - \hat{E}_{\mathcal{F}}\big]$. As is evident from the representation in (6), this variance depends on products of second moments such as the product $\sigma_i^2\sigma_j^2$. While $\hat{\sigma}_i^2$ and $\hat{\sigma}_j^2$ are unbiased for σ_i^2 and σ_j^2, their product is not unbiased, as the estimation error is correlated across the two estimators. Some of this dependence can be removed by leaving both i and j out, but a bias remains as the remaining sample is used in estimating both σ_i^2 and σ_j^2. We therefore propose a leave-*three*-out estimator of the variance product $\sigma_i^2\sigma_j^2$. The product $\boldsymbol{x}_j'\boldsymbol{\beta}\,\boldsymbol{x}_k'\boldsymbol{\beta}\,\sigma_i^2$ appearing in the second component of $\mathbb{V}_0\big[\mathcal{F} - \hat{E}_{\mathcal{F}}\big]$ can similarly be estimated without bias using leave-three-out estimators.

Towards this end, let $\hat{\boldsymbol{\beta}}_{-ijk} = \big(\sum_{\ell\neq i,j,k}\boldsymbol{x}_\ell\boldsymbol{x}_\ell'\big)^{-1}\sum_{\ell\neq i,j,k}\boldsymbol{x}_\ell y_\ell$ denote the OLS estimator of $\boldsymbol{\beta}$ applied to the sample that leaves observations i, j, and k out. Then, define a leave-three-out estimator of σ_i^2 as

$$\hat{\sigma}_{i,-jk}^2 = y_i(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{-ijk}).$$

When j and k are identical, only two observations are left out, and we also write $\hat{\boldsymbol{\beta}}_{-ij}$ and $\hat{\sigma}_{i,-j}^2$. To construct an estimator of $\sigma_i^2\sigma_j^2$, we first write the leave-two-out variance estimator $\hat{\sigma}_{i,-j}^2$ as a weighted sum (see Section 2.7 for details)

$$\hat{\sigma}_{i,-j}^2 = y_i\sum_{k\neq j}\check{M}_{ik,-ij}y_k \qquad \text{where} \qquad \check{M}_{ik,-ij} = \frac{M_{jj}M_{ik} - M_{ij}M_{jk}}{D_{ij}}. \tag{7}$$

Then we multiply each summand above by a leave-three-out variance estimator $\hat{\sigma}_{j,-ik}^2$, which leads to an unbiased estimator of $\sigma_i^2\sigma_j^2$:

$$\widehat{\sigma_i^2\sigma_j^2} = y_i\sum_{k\neq j}\check{M}_{ik,-ij}y_k\cdot\hat{\sigma}_{j,-ik}^2. \tag{8}$$

While this construction appears to treat i and j in an asymmetric fashion, we show to the contrary that (8) is invariant to a permutation of the indices; $\widehat{\sigma_i^2\sigma_j^2} = \widehat{\sigma_j^2\sigma_i^2}$.

To understand why this proposal is unbiased for $\sigma_i^2\sigma_j^2$, it is useful to highlight that $\hat{\sigma}^2_{j,-ik}$ is conditionally independent of (y_i, y_k) and unbiased for σ_j^2, which, when coupled with (7), leads to unbiasedness immediately:

$$\mathbb{E}\big[\,\widehat{\sigma_i^2\sigma_j^2}\,\big] = \sum_{k\neq j}\mathbb{E}\big[y_i\check{M}_{ik,-ij}y_k\big]\cdot\mathbb{E}\big[\hat{\sigma}^2_{j,-ik}\big] = \mathbb{E}\big[\hat{\sigma}^2_{i,-j}\big]\sigma_j^2 = \sigma_i^2\sigma_j^2.$$

An unbiased estimator of the variance expression in (6) that utilizes the variance product estimator in (8) is

$$\hat{V}_{\mathcal{F}} = \sum_{i=1}^{n}\sum_{j\neq i}\big(U_{ij}-V_{ij}^2\big)\cdot\widehat{\sigma_i^2\sigma_j^2} + \sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\neq i}V_{ij}y_j\cdot V_{ik}y_k\cdot\hat{\sigma}^2_{i,-jk}.$$

Note that the product of the $(j=k)$-th terms in the second component generate, for each i, a term not present in (6) and whose non-zero expectation contains $V_{ij}^2\sigma_i^2\sigma_j^2$; hence the use of $U_{ij}-V_{ij}^2$ instead of U_{ij} in the first component.

Remark 1. In the process of establishing the asymptotic validity of the proposed test, we show that the variance estimator $\hat{V}_{\mathcal{F}}$ is close to the null variance $\mathbb{V}_0\big[\mathcal{F}-\hat{E}_{\mathcal{F}}\big]$. In particular, this property implies that the variance estimator is positive with probability approaching one in large samples. However, negative values may still emerge in small samples. In such cases, we propose to replace the variance estimator with an upward biased alternative that uses squared outcomes as estimators of *all* the error variances. This replacement is guaranteed positive, as is detailed in the Appendix, and therefore ensures that the critical value is always defined. Relatedly, Section 4 considers settings where the design matrix may turn rank deficient after leaving certain triples of observations out of the sample. There, we similarly propose to use squared outcomes as estimators of *some* error variances, namely those whose observations cause rank deficiency when left out of the sample.

2.7 Computational remarks

While the previous subsections introduced the location estimator $\hat{E}_{\mathcal{F}}$, variance estimator $\hat{V}_{\mathcal{F}}$, and empirical weights $\hat{\boldsymbol{w}}$ using leave-out estimators of $\boldsymbol{\beta}$, we note here that direct computation of $\hat{\boldsymbol{\beta}}_{-i}$, $\hat{\boldsymbol{\beta}}_{-ij}$, and $\hat{\boldsymbol{\beta}}_{-ijk}$ can be avoided by using the Sherman-Morrison-Woodbury (SMW) identity. Specifically, (4) implies that

$$y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{-i} = \frac{y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}}{M_{ii}},$$

so that computation of $\hat{\boldsymbol{\beta}}_{-i}$ can be avoided when constructing the leave-one-out variance estimator $\hat{\sigma}_i^2 = y_i(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{-i})$. Similarly, it is possible to show that for i and j not equal,

$$y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{-ij} = \frac{M_{jj}(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) - M_{ij}(y_j - \boldsymbol{x}_j'\hat{\boldsymbol{\beta}})}{D_{ij}},$$

which leads to (7), and, for i, j, and k, all of which are different,

$$y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{-ijk} = \frac{(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) - M_{ij}(y_j - \boldsymbol{x}_j'\hat{\boldsymbol{\beta}}_{-jk}) - M_{ik}(y_k - \boldsymbol{x}_k'\hat{\boldsymbol{\beta}}_{-jk})}{D_{ijk}/D_{jk}}.$$

These relationships allow for recursive computation of the leave-out residuals and therefore for simple construction of the variance estimators $\hat{\sigma}_i^2$, $\hat{\sigma}_{i,-j}^2$, and $\hat{\sigma}_{i,-jk}^2$ needed to compute the components of the critical value c_α. In particular, the location estimator $\hat{E}_{\mathcal{F}}$ and empirical weights $\hat{\boldsymbol{w}}$, which require only the leave-one-out residuals, can be computed without explicit loops, by relying instead on elementary matrix operations applied to the matrices containing M_{ij} and B_{ij} as well as the data matrices. Similarly, all doubly indexed objects entering the variance estimator $\hat{V}_{\mathcal{F}}$ can be computed by elementary matrix operations. Those objects are D_{ij}, V_{ij}, U_{ij}, and the leave-two-out residuals. The remaining objects entering $\hat{V}_{\mathcal{F}}$ can be computed by a single loop across i with matrices containing D_{ijk} and leave-three-out residuals renewed at each iteration.

Additionally, the above representations of leave-out residuals demonstrate how M_{ii}^{-1}, D_{ij}^{-1} and D_{ijk}^{-1} enter the critical value, and thus highlight the need for bounding D_{ijk} away from

zero when analyzing the large sample properties of the proposed test.

Remark 2. The quantile $q_{1-\alpha}(\bar{F}_{\hat{\boldsymbol{w}},n-m})$ can easily be constructed by simulating the distribution of the random variable in (1) conditional on the realized value of $\hat{\boldsymbol{w}}$.

3 Asymptotic size and power

This section studies the asymptotic properties of the proposed test. Specifically, we provide a set of regularity conditions under which the test has a correct asymptotic size and non-trivial power against local alternatives. All limits are taken as the sample size n approaches infinity. In studying asymptotic size, we allow for the number of restrictions r and/or number of regressors m to be fixed or diverging with n, but the ordering $r \leq m < n-3$ is maintained throughout the analysis. When studying asymptotic power, we focus on the case of many restrictions, i.e. r diverging with n.

3.1 Asymptotic size

In order to establish asymptotic validity of the proposed test, we impose tail restrictions on the data in addition to a Lindeberg condition that ensures convergence in distribution. When the number of tested restrictions is fixed, the Lindeberg condition implies that the estimator of the tested contrasts $\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{q}$ is asymptotically normal. When the number of restrictions is growing, the Lindeberg condition is weaker and involves only a high-level transformation of the regressors $\sum_{j\neq i} V_{ij}\boldsymbol{x}_j'\boldsymbol{\beta}$, which enters the centered statistic $\mathcal{F} - \hat{E}_{\mathcal{F}}$ as a weight on the i-th error term ε_i. To ensure that the asymptotic distribution of $\mathcal{F} - \hat{E}_{\mathcal{F}}$ does not depend on the unknown distribution of any one error term, we therefore require that no squared $\sum_{j\neq i} V_{ij}\boldsymbol{x}_j'\boldsymbol{\beta}$ dominates the variance $\mathbb{V}_0\big[\mathcal{F} - \hat{E}_{\mathcal{F}}\big]$, which in turn is proportional to r.

Assumption 2. *(i)* $\max_{1\leq i\leq n}\left(\mathbb{E}[\varepsilon_i^4|\boldsymbol{x}_i] + \sigma_i^{-2}\right) = O_p(1)$*, and there exists a sequence* $\{\epsilon_n\}_{n=1}^{\infty}$ *with* $\epsilon_n \to 0$ *such that (ii)* $\epsilon_n^{1/3}\max_{1\leq i\leq n}(\boldsymbol{x}_i'\boldsymbol{\beta})^2 = O_p(1)$ *and (iii) at least one of the following two conditions is satisfied:*

(a) $\max_{1\leq i\leq n} B_{ii} = o_p(\epsilon_n)$,

(b) $\max_{1\leq i\leq n}\left(\sum_{j\neq i}V_{ij}\boldsymbol{x}_j'\boldsymbol{\beta}\right)^2/r=o_p(1)$ *and* $\epsilon_n r\to\infty$.

Part (i) of Assumption 2 limits the thickness of the tails in the error distribution, which is typically required for analysis of OLS estimators (see, e.g., Cattaneo et al., 2018, p.10). Part (ii), which places bounds on $\boldsymbol{x}_i'\boldsymbol{\beta}$, is used to control the variance of the leave-out estimators $\hat{\sigma}_i^2$ and $\hat{\sigma}_{i,-jk}^2$. This condition is used to establish both size control and local power properties, so we stress that it pertains to the actual data generating process, not just the hypothesized value of $\boldsymbol{\beta}$. Note that $\boldsymbol{x}_i'\boldsymbol{\beta}$ may have an unbounded support and the maximum over i may be slowly diverging with n. Part (iii) is a Lindeberg condition as discussed above and can be verified in particular applications of interest. For example, the Appendix shows that $(iii)(b)$ holds in models characterized by group specific regressors.

The next assumption imposes the previously discussed regularity condition that the determinant D_{ijk} is bounded away from zero for any i, j, and k, all of which are different. This condition will be relaxed in Section 4, where such a version of $\hat{V}_{\mathcal{F}}$ is introduced that exists even when leaving two or three observations out leads to rank deficiency of the design.

Assumption 3. $\max_{i\neq j\neq k\neq i}D_{ijk}^{-1}=O_p(1)$.

Under the regularity conditions in Assumptions 2 and 3, the following theorem establishes the asymptotic validity of the proposed testing procedure.

Theorem 3.1. *If Assumptions 1, 2, and 3 hold, then, under* H_0,

$$\lim_{n\to\infty}\mathbb{P}\left(F>c_\alpha\right)=\alpha.$$

3.2 Asymptotic power

To describe the power of the proposed test, we introduce a drifting sequence of local alternatives indexed by a deviation $\boldsymbol{\delta}$ from the null times $(\boldsymbol{R}\boldsymbol{S}_{xx}^{-1}\boldsymbol{R}')^{1/2}$, which specifies the precision the tested linear restrictions can be estimated with in the given sample. Thus, we consider

alternatives of the form

$$H_\delta \ : \ \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{q} + (\boldsymbol{R}\boldsymbol{S}_{xx}^{-1}\boldsymbol{R}')^{1/2} \cdot \boldsymbol{\delta}, \tag{9}$$

for $\boldsymbol{\delta} \in \mathbb{R}^r$ satisfying the limiting condition

$$\lim_{n,r\to\infty} \frac{\|\boldsymbol{\delta}\|}{r^{1/4}} = \Delta_\delta \in [0, \infty].$$

Below we show that the power of the test is monotone in Δ_δ, with power equal to size when $\Delta_\delta = 0$ and power equal to one when $\Delta_\delta = \infty$.

The role of $(\boldsymbol{R}\boldsymbol{S}_{xx}^{-1}\boldsymbol{R}')^{1/2}$ in indexing the local alternatives is analogues to that of $n^{-1/2}$ often used in parametric problems. However, in settings with many regressors some linear restrictions may be estimated at rates that are substantially lower than the standard parametric one. Therefore, we index the deviations from the null by the actual rate of $(\boldsymbol{R}\boldsymbol{S}_{xx}^{-1}\boldsymbol{R}')^{1/2}$ instead of $n^{-1/2}$.

The alternative is additionally indexed by $\boldsymbol{\delta}$, which in standard parametric problems is typically fixed. However, fixed $\boldsymbol{\delta}$ is less natural here, as the dimension of $\boldsymbol{\delta}$ increases with sample size. Instead, we fix the limit of its Euclidean norm when scaled by $r^{1/4}$. This approach allows us to discuss different types of alternatives and how the numerosity of the tested restrictions affects the test's ability to detect deviations from the null. Specifically, note that when the deviation $\boldsymbol{\delta}$ is sparse, i.e., only a bounded number of its entries are non-zero, then the test has a non-trivial power against alternatives that diverge at a rate that is $r^{1/4}$ lower than when only a fixed number of restrictions is being tested. This observation highlights the cost for the power of including many *irrelevant* restrictions in the hypothesis. On the other hand, if $\boldsymbol{\delta}$ is dense, e.g., with all entries bounded away from zero, then the test can detect local deviations that on average shrink at a rate that is $r^{1/4}$ greater than the usual. This means that if the tested restrictions can be estimated at the parametric rate and they are all *relevant*, then the test can detect deviations from the null of order $n^{-1/2}r^{-1/4}$.

The following theorem states the asymptotic power under sequences of local alternatives

of the form given in (9) and discussed above.

Theorem 3.2. *If Assumptions 1, 2, and 3 hold, then, under* H_δ,

$$\lim_{n,r\to\infty} \mathbb{P}\left(F > c_\alpha\right) - \Phi\left(\Phi^{-1}\left(\alpha\right) + \Delta_\delta^2\left(\mathbb{V}_0\big[\mathcal{F} - \hat{E}_\mathcal{F}\big]/r\right)^{-1/2}\right) = 0,$$

where Φ *denotes the cumulative distribution function of the standard normal and* $\Phi(\infty) = 1$.

Remark 3. It is instructive to compare the power curve documented in Theorem 3.2 with the asymptotic power curve of the exact F test when both tests are valid. When the individual error terms are homoskedastic normal with variance σ^2, the asymptotic power of the exact F test is the limit of (Anatolyev, 2012)

$$\Phi\left(\Phi^{-1}\left(\alpha\right) + \Delta_\delta^2\left(2\sigma^4 + 2\sigma^4 r/(n-m)\right)^{-1/2}\right).$$

Thus, the relative asymptotic power of the proposed LO test and the exact F test is determined by the limiting ratio of $r^{-1}\mathbb{V}_0\big[\mathcal{F} - \hat{E}_\mathcal{F}\big]$ to $2\sigma^4(1 + r/(n-m))$. The Appendix shows that this ratio approaches one in large samples if the number of tested restriction is small relative to the sample size or if the limiting variability of B_{ii}/M_{ii} is small (Kline et al., 2020, calls this a balanced design). When neither of these conditions holds, the proposed test will, in general, have a slightly lower power than the exact F test, which we also document in the simulations in Section 5.

Remark 4. The order of the numerosity of alternatives that can be detected with the proposed test is optimal in the minimax sense when the alternatives are moderately sparse to dense, i.e., when $O(\sqrt{r})$ or more of the tested restrictions are violated (Arias-Castro et al., 2011). However, if the alternative is strongly sparse so that at most $o(\sqrt{r})$ tested restrictions are violated, a higher power can be achieved by tests that redirect their power towards those alternatives. Such tests typically focus their attention on a few largest t statistics (i.e., smallest p values) and are often described as multiple comparison procedures (Donoho and Jin, 2004; Romano et al., 2010). While such tests can control size when the error terms are homoskedastic normal, it is not clear whether they can do so in the current semiparametric

framework with an unspecified error distribution. The issue is that the size control for multiple comparisons relies on knowing the (normal or t) distributions of individual t statistics, but in the current framework with many regressors those distributions are not necessarily known (even asymptotically).

4 If leave-three-out fails

This section extends the definition of the critical value c_α to settings where the design matrix may turn rank deficient after leaving certain pairs or triples of observations out of the sample. When Assumption 1 fails in this way, $\hat{E}_{\mathcal{F}}$ is still an unbiased estimator of $\mathbb{E}_0[\mathcal{F}]$, but the unbiased variance estimator introduced in Section 2.6 does not exist. For this reason, we propose an adjustment to the variance estimator that introduces a positive bias for pairs of observations where we are unable to construct an unbiased estimator of the variance product $\sigma_i^2\sigma_j^2$ and for triples of observations where we are unable to construct an unbiased estimator of $\boldsymbol{x}_j'\boldsymbol{\beta}\,\boldsymbol{x}_k'\boldsymbol{\beta}\,\sigma_i^2$. This introduction of a positive bias to the variance estimator ensures asymptotic size control, even when Assumption 1 fails.

Since this section considers a setup where Assumption 1 may fail, we introduce a weaker version of the assumption, which only imposes the full rank of the design matrix after dropping any one observation.

Assumption 1'. $\sum_{j\neq i}\boldsymbol{x}_j\boldsymbol{x}_j'$ *is invertible for every* $i \in \{1,\dots,n\}$.

One can always satisfy this assumption by appropriately pruning the sample, the model, and the hypothesis of interest. For example, if $\boldsymbol{S}_{xx}$ does not have full rank, then one can remove unidentified parameters from both the model and hypothesis of interest, and proceed by testing the subset of restrictions in H_0 that are identified by the sample. Similarly, if $\sum_{j\neq i}\boldsymbol{x}_j\boldsymbol{x}_j'$ does not have full rank for some observation i, then there is a parameter in the model which is identified only by this observation. Therefore, one can proceed as in the case of rank deficiency of $\boldsymbol{S}_{xx}$, by dropping observation i from the sample and by removing the parameter that determines the mean of this observation from the model and null hypothesis.

When doing this for any observation i such that $\sum_{j\neq i} \boldsymbol{x}_j\boldsymbol{x}_j'$ is non-invertible, one obtains a sample that satisfies Assumption 1' and can be used to test the restrictions in H_0 that are identified by this *leave-one-out sample*.

4.1 Variance estimator

When Assumption 1 fails, some of the unbiased estimators $\hat{\sigma}^2_{i,-jk}$ and $\widehat{\sigma_i^2\sigma_j^2}$ cease to exist. For such cases, the variance estimator $\hat{V}_{\mathcal{F}}$ utilizes replacements that are either also unbiased or positively biased, depending on the *cause* of the failure. Assumption 1 fails if $D_{ijk} = 0$ for some triple of observations, and we say that this failure of full rank is *caused by* i if $D_{jk} > 0$ or $D_{ij}D_{ik} = 0$, i.e., if the design retains full rank when only observations j and k are left out or if leaving out observations (i, j) or (i, k) leads to rank deficiency. Our replacement for $\hat{\sigma}^2_{i,-jk}$ is biased when i causes $D_{ijk} = 0$, while the replacement for $\widehat{\sigma_i^2\sigma_j^2}$ is biased when both i and j cause $D_{ijk} = 0$ for some k.

To introduce the replacement for $\hat{\sigma}^2_{i,-jk}$, we consider the case when it does not exist, or equivalently, when $D_{ijk} = 0$. If i causes this leave-three-out failure, then our replacement is the upward biased estimator y_i^2. When this failure of leave-three-out is not caused by i, the leave-two-out estimators $\hat{\sigma}^2_{i,-j}$ and $\hat{\sigma}^2_{i,-k}$ are equal and independent of both y_j and y_k (as shown in the Appendix). These properties imply that $y_j y_k \hat{\sigma}^2_{i,-j}$ is an unbiased estimator of $\boldsymbol{x}_j'\boldsymbol{\beta}\,\boldsymbol{x}_k'\boldsymbol{\beta}\,\sigma_i^2$, and we therefore use $\hat{\sigma}^2_{i,-j}$ as a replacement for $\hat{\sigma}^2_{i,-jk}$. To summarize, we let

$$\bar{\sigma}^2_{i,-jk} = \begin{cases} \hat{\sigma}^2_{i,-jk}, & \text{if } D_{ijk} > 0, \\ \hat{\sigma}^2_{i,-j}, & \text{if } D_{jk} = 0 \text{ and } D_{ij}D_{ik} > 0, \\ y_i^2, & \text{otherwise.} \end{cases}$$

When j is equal to k, we consider pairs of observations, and the definition only involves the last two lines since $D_{ijj} = 0$. In this case, we also write $\bar{\sigma}^2_{i,-j}$ for $\bar{\sigma}^2_{i,-jj}$.

For the replacement of $\widehat{\sigma_i^2\sigma_j^2} = y_i \sum_{k\neq j} \check{M}_{ik,-ij} y_k \cdot \hat{\sigma}^2_{j,-ik}$, we similarly consider the case where this estimator does not exist, i.e., where $D_{ijk} = 0$ for a k not equal to i or j. When

any such rank deficiency is caused by both i and j, we rely on the upward biased replacement $y_i^2 \bar{\sigma}^2_{j,-i}$. When none of the leave-three-out failures are caused by both i and j, the replacement uses $\bar{\sigma}^2_{i,-jk}$ in place of $\hat{\sigma}^2_{i,-jk}$. To summarize, we define

$$\overline{\sigma_i^2 \sigma_j^2} = \begin{cases} y_i \sum_{k \neq j} \check{M}_{ik,-ij} y_k \cdot \bar{\sigma}^2_{j,-ik}, & \text{if } D_{ij} > 0 \text{ and } (D_{ijk} > 0 \text{ or } D_{ik} D_{jk} = 0 \text{ for all } k), \\ y_i^2 \bar{\sigma}^2_{j,-i}, & \text{otherwise.} \end{cases}$$

This estimator is unbiased for $\sigma_i^2 \sigma_j^2$ when none of the leave-three-out failures are caused by both i and j, i.e., when the first line of the definition applies. Unbiasedness holds because the presence of a bias in $\bar{\sigma}^2_{j,-ik}$ implies that j is causing the leave-three-out failure. Therefore, i cannot be the cause, which yields that $\hat{\sigma}^2_{i,-j}$ is independent of y_k, or equivalently, that $\check{M}_{ik,-ij} = 0$.

Now, we describe how these replacement estimators enter the variance estimator $\hat{V}_{\mathcal{F}}$. When $\overline{\sigma_i^2 \sigma_j^2}$ or $\bar{\sigma}^2_{i,-jk}$ are biased and would enter the variance estimator with a negative weight, we remove these terms, as they would otherwise introduce a negative bias. For $\overline{\sigma_i^2 \sigma_j^2}$, the weight is $U_{ij} - V_{ij}^2$, so a biased variance product estimator is removed when $U_{ij} - V_{ij}^2 < 0$. For $\bar{\sigma}^2_{i,-jk}$, the weight is $V_{ij} y_j \cdot V_{ik} y_k$, but $\bar{\sigma}^2_{i,-jk}$ does not depend on j and k when it is biased, so we sum these weights across all such j and k, and we remove the term if this sum is negative.

The following variance estimator extends the definition of $\hat{V}_{\mathcal{F}}$ to settings where leave-three-out may fail:

$$\hat{V}_{\mathcal{F}} = \sum_{i=1}^{n} \sum_{j \neq i} \left(U_{ij} - V_{ij}^2 \right) \cdot G_{ij} \cdot \overline{\sigma_i^2 \sigma_j^2} + \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k \neq i} V_{ij} y_j \cdot V_{ik} y_k \cdot G_{i,-jk} \cdot \bar{\sigma}^2_{i,-jk},$$

where the indicators G_{ij} and $G_{i,-jk}$ remove biased estimators with negative weights:

$$G_{ij} = \begin{cases} 0, & \text{if } \overline{\sigma_i^2 \sigma_j^2} = y_i^2 \bar{\sigma}_{j,-i}^2 \text{ and } U_{ij} - V_{ij}^2 < 0, \\ 1, & \text{otherwise,} \end{cases}$$

$$G_{i,-jk} = \begin{cases} 0, & \text{if } \bar{\sigma}_{i,-jk}^2 = y_i^2 \text{ and } \sum_{j \neq i} \sum_{k \neq i} V_{ij} y_j \cdot V_{ik} y_k \cdot \mathbf{1}\{\bar{\sigma}_{i,-jk}^2 = y_i^2\} < 0, \\ 1, & \text{otherwise.} \end{cases}$$

4.2 Asymptotic size

In order to establish that the proposed test controls asymptotic size when there are some failures of leave-three-out, we replace the regularity condition in Assumption 3 with an analogous version that allows for some of the determinants D_{ij} and D_{ijk} to be zero. Otherwise, the role of Assumption 3' below is the same as Assumption 3 in that it rules out denominators that are arbitrarily close to zero.

Assumption 3'. *(i)* $\max_{1 \leq i \leq n} M_{ii}^{-1} = O_p(1)$, *and (ii)* $\max_{i,j: D_{ij} \neq 0} D_{ij}^{-1} + \max_{i,j,k: D_{ijk} \neq 0} D_{ijk}^{-1} = O_p(1)$.

When computing $\hat{V}_{\mathcal{F}}$, one must account for machine zero imperfections while comparing D_{ij} and D_{ijk} with zero in the definitions of $\bar{\sigma}_{i,-jk}^2$ and $\overline{\sigma_i^2 \sigma_j^2}$. Such imperfections are typically of order 10^{-15}; however, we propose to compare D_{ij} to 10^{-4} and D_{ijk} to 10^{-6}. Doing so will replace any potential case of a small denominator with an upward biased alternative and ensures that Assumption 3'(ii) is automatically satisfied.

The following theorem establishes the asymptotic validity of the proposed leave-out test in settings where Assumption 1 fails. The theorem pertains to a nominal size below 0.31, as the upward biased variance estimator may not ensure validity in cases where a nominal size above 0.31 is desired. This happens because the quantile $q_{1-\alpha}(\bar{F}_{\hat{\boldsymbol{w}},n-m})$ may fall below 1 when α is greater than 0.31.

Theorem 4.1. *If* $\alpha \in (0, 0.31]$ *and Assumptions 1', 2, and 3' hold, then, under* H_0,

$$\limsup_{n \to \infty} \mathbb{P}\left(F > c_\alpha\right) \leq \alpha.$$

An important difference between this result and that of Theorem 3.1 is that the asymptotic size may be smaller than desired, which can happen when leave-three-out fails for a large fraction of possible triples. When such conservatism materializes, there will be a corresponding loss in power relative to the result in Theorem 3.2. Otherwise, the power properties are analogous to those reported in Theorem 3.2 and we therefore omit a formal result.

Remark 5. Before turning to a study of the finite sample performance of the proposed test, we describe an adjustment to the test which is based on finite sample considerations. This adjustment is to rely on demeaned outcome variables in the definitions of $\hat{E}_{\mathcal{F}}$, $\hat{V}_{\mathcal{F}}$, and $\hat{\boldsymbol{w}}$. The benefit of relying on demeaned outcomes is that it makes the critical value invariant to the location of the outcomes. On the other hand, this adjustment removes the exact unbiasedness used to motivate the estimators of $\mathbb{E}_0[\mathcal{F}]$ and $\mathbb{V}_0\big[\mathcal{F} - \hat{E}_{\mathcal{F}}\big]$. However, one can show that the biases introduced by demeaning vanish at a rate that ensures asymptotic validity. Therefore, we deem the gained location invariance sufficiently desirable that we are willing to introduce a small finite sample bias to achieve it. We refer to the Appendix for exact mathematical details but note that this adjustment is used in the following simulations.

5 Simulation evidence

This section documents some finite sample properties of the proposed leave-out (LO) test and compares its performance with conventional tests that are likely to be used by a researcher in the present context. These benchmark tests are the following:

1. The exact F test that uses critical values from the F distribution to reject when $F > q_{1-\alpha}(F_{r,n-m})$. This test has actual size equal to nominal size in finite samples under conditional homoskedastic normal errors for any number of regressors and restrictions. It is also asymptotically valid with conditional homoskedasticity and non-normality under certain homogeneity conditions on the regressors.

2. Three different Wald tests that reject when a heteroskedasticity-robust Wald statistic

exceeds the $(1-\alpha)$-th quantile of a χ^2_r distribution, i.e., when $W > q_{1-\alpha}(\chi^2_r)$ for

$$W = \big(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{q}\big)' \Big(\boldsymbol{R}\boldsymbol{S}_{xx}^{-1} \Big(\sum\nolimits_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \tilde{\sigma}_i^2\Big) \boldsymbol{S}_{xx}^{-1} \boldsymbol{R}'\Big)^{-1} \big(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{q}\big)$$

and particular choices of the error variance estimators $\{\tilde{\sigma}_i^2\}_{i=1}^n$. We consider a degrees-of-freedom corrected individual error variance estimator $\tilde{\sigma}_i^2 = (y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})^2 n/(n-m)$ which is referred to as W_1 (MacKinnon, 2013), a leave-one-out version with $\tilde{\sigma}_i^2 = \hat{\sigma}_i^2$ that we call W_{L} (Kline et al., 2020), and the version considered in Cattaneo et al. (2018) which we refer to as W_{K}. Asymptotically, the two Wald tests, W_{L} and W_{K}, are valid with many regressors under arbitrary heteroskedasticity but not necessarily with many restrictions, while W_1 is valid with few regressors and few restrictions under arbitrary heteroskedasticity.

5.1 Simulation design

The simulation setup borrows elements of MacKinnon (2013) and adapts it to the case of many regressors as in Richard (2019) but with richer heterogeneity in the design. The outcome equation is

$$y_i = \beta_1 + \sum_{k=2}^{m} \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where data is drawn *i.i.d.* across i. Following MacKinnon (2013), the sample sizes take the values 80, 160, 320, 640, and 1280. The number of unknown coefficients is $m = 0.8n$ throughout to demonstrate the validity of the proposed test even with very many regressors. The null restricts the values of the last r coefficients using $\boldsymbol{R} = \big[\boldsymbol{0}_{r\times(m-r)}, \boldsymbol{I}_r\big]$. We consider both a design that contains only continuous regressors and a *mixed* one that also includes some discrete regressors.

In the continuous design, the regressors $x_{i2}, \ldots, x_{im}$ are products of independent standard log-normal random variables and a common multiplicative mean-unity factor drawn independently from a shifted standard uniform distribution, i.e., $0.5 + u_i$ where u_i is standard uniform. This common factor induces dependence among the regressors and rich heterogene-

ity in the statistical leverages of individual observations. For this design, we consider $r = 3$ and $r = 0.6n$.

When also including discrete regressors, we let $x_{i2}, \ldots, x_{i,m-r}$ be as above and let the last r regressors be group dummies. This mixed design corresponds to random assignment into $r + 1$ groups with the last group effect removed due to the presence of an intercept in the model. The assigned group number is the integer ceiling of $(r + 1)(u_i + u_i^2)/2$, where u_i is the multiplicative factor used to generate dependence among the continuous regressors. By reusing u_i we maintain dependence between all regressors, and by using a nonlinear transformation of u_i we induce systematic variability among the $r + 1$ expected group sizes. We let $r = 0.15n$, which leads the expected group sizes to vary between 4 and 13 with an average group size of about 6.5. The null corresponds to a hypothesis of equality of means across all groups.

Each regression error is a product of a standard normal random variable and an individual specific standard deviation σ_i. The standard deviation is generated by

$$\sigma_i = z_\zeta \left(1 + s_i\right)^\zeta, \quad i = 1, \ldots, n,$$

where $s_i > 0$ depends on the design and the multiplier z_ζ is such that the mean of σ_i^2 is unity. The parameter $\zeta \in [0, 2]$ indexes the strength of heteroskedasticity, with $\zeta = 0$ corresponding to homoskedasticity. We consider only the two extreme cases of $\zeta \in \{0, 2\}$. In the continuous design, we let $s_i = \sum_{k=2}^{m} x_{ik}$, and in the mixed design, $s_i = \sum_{k=2}^{m-r} x_{ik} + z_u u_i$. The factor $z_u = 2r \exp(1/2)$ ensures that s_i has the same mean in both designs.

Under the null, the coefficients on the continuous regressors are all equal to ϱ, where ϱ is such that the coefficient of determination, R^2, equals 0.16. The coefficients on the included group dummies are zero, which correspond to the null of equality across all groups. The intercept is chosen such that the mean of the outcomes is unity. For the continuous design this yields an intercept of $1-(m-1)\varrho \exp(1/2)$, while the intercept is $1-(m-r-1)\varrho \exp(1/2)$ in the mixed design. With these parameter values, the null is $(\beta_{m-r}, \ldots, \beta_r)' = \boldsymbol{q}$, where $\boldsymbol{q} = (\varrho, \ldots, \varrho)' \in \mathbb{R}^r$ in the continuous design, and $\boldsymbol{q} = (0, \ldots, 0)' \in \mathbb{R}^r$ in the mixed design.

To document power properties, we consider both a sparse and dense deviations from the null, and focus on the settings where r is proportional to n. In parallel to the theoretical power analysis in Section 3, we consider deviations for the last r coefficients that are parameterized using

$$(\beta_{m-r}, \ldots, \beta_m)' = \boldsymbol{q} + \left(\boldsymbol{R}\mathbb{E}[\boldsymbol{S}_{xx}]^{-1}\boldsymbol{R}'\right)^{1/2}\boldsymbol{\delta},$$

where we use the lower triangular square-root matrix. This choice of square-root implies that the alternative is sparse when only the last few entries of $\boldsymbol{\delta}$ are non-zero. As shown in Section 3, asymptotic power is governed by the norm of $\boldsymbol{\delta}$ over $r^{1/4}$, but whether an alternative is fixed or local, additionally depends on the rate at which the tested coefficients are estimated. This rate is governed by $\mathbb{E}[\boldsymbol{S}_{xx}]$, which is reported in the Appendix.

In the continuous design, the tested coefficients are estimated at the standard parametric rate of $n^{-1/2}$. To specify a fixed sparse alternative we therefore use $\boldsymbol{\delta} = 0.5n^{1/2}(0, \ldots, 0, 1)' \in \mathbb{R}^r$, for which β_m differs from the null value by approximately 0.2 (here and hereafter, the scaling is chosen so that the power is bounded away from the size and away from unity for the sample sizes we consider). Since the norm of $\boldsymbol{\delta}$ grows faster than $r^{1/4}$, the power will be an increasing function of the sample size. For the dense alternative, we consider instead $\boldsymbol{\delta} = 0.5n^{1/2}r^{-1/2}\boldsymbol{\iota}_r$ where $\boldsymbol{\iota}_r = (1, \ldots, 1)' \in \mathbb{R}^r$, for which all deviations between the tested coefficients and ϱ shrink at the standard parametric rate of $n^{-1/2}$. Here, power is again increasing in the sample size due to numerous deviations from the null.

In the mixed design, the group effects are not estimated consistently as the group sizes are bounded. A possible fixed sparse alternative is then $\boldsymbol{\delta} = (0, \ldots, 0, 6)' \in \mathbb{R}^r$, for which β_m differs from the null value of zero by roughly 3. In contrast to the continuous design, the power will decrease with sample size as the precision, with which β_m can be estimated, does not increase with n. For the dense alternative, we use $\boldsymbol{\delta} = 1.5\boldsymbol{\iota}_r$, which corresponds to a fixed alternative for *every* tested coefficient. Here, the power will be increasing in n due to the numerosity of deviations.

5.2 Simulation results

We present rejection rates based on 10000 Monte-Carlo replications and consider tests with nominal sizes of 1%, 5% and 10%. Furthermore, we report the frequency with which the proposed variance estimate $\hat{V}_{\mathcal{F}}$ is negative and therefore replaced by the upward biased and positive alternative introduced in Remark 1. For the design that includes discrete regressors, we also report the average fraction of observations that cause a failure of leave-three-out full rank, and for which we therefore rely on an upward biased estimator of the corresponding error variance. For all sample sizes, this fraction is around 7% in the mixed design, which corresponds to the percentage of observations that belong to groups of size 2 or 3. The fraction is zero in the design that only involves continuous regressors.

Table 1 contains the actual rejection rates under the null for both the continuous and mixed designs. In settings with many regressors and restrictions, the considered versions of the "heteroskedasticity-robust" Wald test fail to control size irrespective of the design, presence of heteroskedasticity, and nominal size. The failure of the conventional Wald test, W_1, is spectacular, with type I error rates close to one for the continuous design, but the two versions that are robust to many regressors, W_K and W_L, also exhibit size well above the nominal level. With few restrictions, the Wald tests show a more moderate inability to match actual size with nominal size, and the table suggests that the leave-one-out version, W_L, can control size in samples that are somewhat larger than considered here. Under homoskedasticity, the table reports that the exact F test indeed has exact size. However, in the heteroskedastic environments with many restrictions the exact F test is oversized with a type I error rate that approaches unity as the sample size increases.

By contrast, the proposed leave-out test exhibits nearly flawless size control as it is oversized by at most one percent across nearly all designs, nominal sizes, and whether heteroskedasticity is present or not. In the smallest sample for the continuous design, the test is somewhat conservative, presumably due to the relatively high rate of negative variance estimates (20% with homoskedasticity and 13% with heteroskedasticty) that are replaced by a strongly upward biased alternative. This rate diminishes quickly with sample size, and

Table 1: Empirical size (in percent)

Nominal size		1%					5%					10%					
Test		LO	EF	W_1	W_K	W_L	LO	EF	W_1	W_K	W_L	LO	EF	W_1	W_K	W_L	$\%\hat{V}_{\mathcal{F}} < 0$
Homoskedasticity																	
Continuous design																	
$n = 80$	$r = 3$	2	1	4	14	11	7	5	10	19	18	12	10	15	23	23	8.7
$n = 160$	$r = 3$	1	1	2	14	7	6	5	6	19	14	12	10	10	23	20	2.0
$n = 320$	$r = 3$	1	1	1	15	4	6	5	4	22	10	11	10	8	27	15	0.6
$n = 640$	$r = 3$	1	1	1	15	2	6	5	3	23	8	11	10	7	29	13	0.1
$n = 1280$	$r = 3$	1	1	1	9	2	6	5	3	17	7	11	10	7	24	12	0.0
$n = 80$	$r = 48$	1	1	99	54	16	3	5	100	54	18	7	10	100	54	19	19.8
$n = 160$	$r = 96$	2	1	100	54	20	5	5	100	54	21	10	10	100	54	22	6.4
$n = 320$	$r = 192$	2	1	100	53	20	6	5	100	54	21	11	10	100	54	21	1.5
$n = 640$	$r = 384$	1	1	100	54	22	6	5	100	54	23	11	10	100	54	23	0.3
$n = 1280$	$r = 768$	1	1	100	53	22	5	5	100	53	23	11	10	100	53	23	0.0
Mixed design																	
$n = 80$	$r = 12$	2	1	21	25	19	7	5	33	29	23	12	10	41	31	26	4.8
$n = 160$	$r = 24$	1	1	25	27	23	6	5	38	30	27	12	10	47	31	29	0.4
$n = 320$	$r = 48$	1	1	33	29	29	5	5	49	31	32	11	10	58	32	34	0.1
$n = 640$	$r = 96$	1	1	47	31	32	5	5	64	33	35	11	10	72	33	36	0.0
$n = 1280$	$r = 192$	1	1	69	31	35	5	5	82	33	37	10	10	87	33	38	
Heteroskedasticity																	
Continuous design																	
$n = 80$	$r = 3$	2	3	7	17	10	6	10	15	22	16	11	17	22	26	21	10.3
$n = 160$	$r = 3$	2	2	4	14	7	6	8	10	20	14	12	15	16	24	19	2.3
$n = 320$	$r = 3$	1	2	2	16	4	6	8	8	23	10	12	15	14	28	16	0.6
$n = 640$	$r = 3$	1	2	2	15	2	6	8	7	23	8	11	14	12	29	13	0.3
$n = 1280$	$r = 3$	1	2	1	9	2	5	8	6	17	6	11	15	12	24	12	0.0
$n = 80$	$r = 48$	1	22	100	52	13	5	47	100	52	15	9	61	100	52	16	12.8
$n = 160$	$r = 96$	1	32	100	50	15	5	61	100	50	16	11	75	100	51	17	3.7
$n = 320$	$r = 192$	1	56	100	49	17	6	81	100	49	18	11	90	100	49	19	0.9
$n = 640$	$r = 384$	1	86	100	49	18	5	96	100	49	19	11	99	100	49	20	0.1
$n = 1280$	$r = 768$	1	99	100	48	19	5	100	100	48	19	11	100	100	48	20	0.0
Mixed design																	
$n = 80$	$r = 12$	1	5	38	27	14	6	17	51	30	18	12	26	58	32	21	4.8
$n = 160$	$r = 24$	1	9	49	28	19	6	24	63	30	23	13	35	71	31	25	0.5
$n = 320$	$r = 48$	1	14	67	29	22	5	33	80	31	25	11	46	85	32	27	0.1
$n = 640$	$r = 96$	1	28	89	31	26	5	52	95	32	29	11	65	97	33	30	0.0
$n = 1280$	$r = 192$	1	52	99	33	30	5	75	100	34	31	10	84	100	34	32	0.0

NOTE: LO: leave-out test, EF: exact F test, W_1: heteroskedastic Wald test with degrees-of-freedom correction, W_K: heteroskedastic Wald test with Cattaneo et al. (2018) correction, W_L: heteroskedastic Wald test with Kline et al. (2020) correction; $\%\hat{V}_{\mathcal{F}} < 0$: fraction of negative variance estimates for LO (in percent). Results from 10000 Monte-Carlo replications.

Table 2: Empirical power (in percent) corresponding to 5% and 10% size

		Homoskedasticity								Heteroskedasticity			
Deviation		Sparse				Dense				Sparse		Dense	
Nominal size		5%		10%		5%		10%		5%	10%	5%	10%
Test		LO	EF	LO	EF	LO	EF	LO	EF	LO	LO	LO	LO
Continuous design													
$n = 80$	$r = 48$	6	15	12	25	5	15	10	25	10	18	7	14
$n = 160$	$r = 96$	16	23	26	34	12	21	22	34	20	32	17	29
$n = 320$	$r = 192$	29	35	43	48	26	36	39	51	31	45	29	44
$n = 640$	$r = 384$	49	55	63	69	44	57	58	71	52	66	49	64
$n = 1280$	$r = 768$	74	80	84	88	68	84	81	92	76	86	74	85
Mixed design													
$n = 80$	$r = 12$	18	23	30	36	17	19	29	30	24	38	23	37
$n = 160$	$r = 24$	18	18	29	28	27	28	41	41	24	35	34	49
$n = 320$	$r = 48$	13	13	22	22	40	42	54	56	16	27	48	64
$n = 640$	$r = 96$	10	10	18	18	60	65	73	77	11	20	70	82
$n = 1280$	$r = 192$	8	8	16	15	87	91	94	95	9	17	92	97

NOTE: LO: leave-out test, EF: exact F test. Results from 10000 Monte-Carlo replications.

the fraction of negative variance estimates is already essentially zero in samples with 640 observations and 512 regressors. In the mixed design, negative variance estimates are even less prevalent, potentially due to the fact that the test uses some upward biased variance estimators for 7% of observations. Perhaps somewhat surprisingly, having 7% of observations causing failure of leave-three-out is not sufficient to bring about any discernible conservativeness in the leave-out test for this design.

Table 2 contains simulated rejection rates for the continuous and mixed designs under alternatives where the parameters deviate from their null values in one of two ways – either one tested coefficient deviates (sparse) or all tested coefficients deviate (dense). The table reports these power figures for tests with a nominal size of 5% and 10% that also control the size well, i.e., the LO and exact F tests under homoskedasticity and the LO test under heteroskedasticity.

For the continuous design, the power of the tests increases from slightly above nominal size to somewhat below unity as the number of observations increases from 80 to 1280. This pattern largely holds irrespective of the type of deviation and presence of heteroskedasticity,

although the LO test is a bit more responsive to sparse deviations than to dense ones. Along this stretch of the power curve, the LO test exhibits a power loss that varies between 4 and 16 percentage points when compared to the exact F test, and in relative terms, this gap in power shrinks as the sample size grows. Given that the number of tested restrictions in this setting is above half of the sample size, we conjecture that these figures are towards the high end of the power loss that a typical practitioner would incur in order to be robust with respect to heteroskedasticity.

In the mixed design, the fixed dense alternative exhibits similar power figures as in the continuous design, while the fixed sparse deviation generates a power function that decreases with sample size. The reason for the latter is, as discussed in the previous subsection, that the deviating group effect is not estimated more precisely as additional groups are added to the data. Upon comparison of the LO and exact F tests, we see that the differences in the power figures are only 0–7 percentage points. In light of Remark 3, which explains that there is no power difference between the LO and exact F tests when r/n is small, it is natural to attribute this almost non-existant power loss to the fact that there are four times fewer tested restrictions in this mixed design than in the continuous one.

6 Concluding remarks

This paper develops an inference method for use in a linear regression with conditional heteroskedasticity where the objective is to test a hypothesis that imposes many linear restrictions on the regression coefficients. The proposed test rejects the null hypothesis if the conventional F statistic exceeds a linearly transformed quantile from the F-bar distribution. The central challenges for construction of the test is estimation of individual error variances and their products, which requires new ideas when the number of regressors is large. We overcome these challenges by using the idea of leaving up to three observations out when estimating individual error variances and their products. In some samples the variance estimate used for rescaling of the critical value may either be negative or cease to exist due to the presence of many discrete regressors. For both of these issues, we propose an automatic

adjustment that relies on intentionally upward biased estimators which in turn leaves the resulting test somewhat conservative. Simulation experiments show that the test controls size in small samples, even in strongly heteroskedastic environments, and only exhibits very limited adjustment-induced conservativeness. The simulations additionally illustrate good power properties that signal a manageable cost in power from relying on a test that is robust to heteroskedasticity and many restrictions.

Bootstrapping and closely related resampling methods are often advocated as automatic approaches for the construction of critical values. However, in the context of linear regression with proportionality between the number of regressors and sample size, multiple papers (Bickel and Freedman, 1983; El Karoui and Purdom, 2018; Cattaneo et al., 2018) demonstrate the invalidity of standard bootstrap schemes even when inferences are made on a single regression coefficient. Under an additional assumption of homoskedasticity and further restrictions on the design, El Karoui and Purdom (2018) and Richard (2019) show that various (problem-specific) corrections to bootstrap methods can restore validity. We leave it to future research to determine whether bootstrap or other resampling methods can be corrected to ensure validity in our context of a heteroskedastic regression model with many regressors and tested restrictions.

References

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*(2), 251–333.

Akritas, M. G. and N. Papadatos (2004). Heteroscedastic one-way ANOVA and lack-of-fit tests. *Journal of American Statistical Association 99*(466), 368–382.

Anatolyev, S. (2012). Inference in regression models with many regressors. *Journal of Econometrics 170*(2), 368–382.

Anatolyev, S. (2013). Instrumental variables estimation and inference in the presence of many exogenous regressors. *Econometrics Journal 16*(1), 27–72.

Anatolyev, S. (2018). Almost unbiased variance estimation in linear regressions with many covariates. *Economics Letters 169*, 20–23.

Anatolyev, S. (2019). Many instruments and/or regressors: a friendly guide. *Journal of Economic Surveys 33*(2), 689–726.

Anatolyev, S. and M. Sølvsten (2020). `manyRegressors`: R package for inference in models with heteroskedasticity and many regressors. https://github.com/mikkelsoelvsten/manyRegressors.

Arias-Castro, E., E. J. Candès, and Y. Plan (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Annals of Statistics 39*(5), 2533–2556.

Berndt, E. R. and N. E. Savin (1977). Conflict among criteria for testing hypotheses in the multivariate linear regression model. *Econometrica 45*(5), 1263–1277.

Bickel, P. J. and D. A. Freedman (1983). Bootstrapping regression models with many parameters. In *A festschrift for Erich L. Lehmann*, pp. 28–48. CRC Press.

Calhoun, G. (2011). Hypothesis testing in linear regression when k/n is large. *Journal of Econometrics 165*(2), 163–174.

Card, D., A. R. Cardoso, J. Heining, and P. Kline (2018). Firms and labor market inequality: Evidence and some theory. *Journal of Labor Economics 36*(S1), S13–S70.

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

Chao, J. C., J. A. Hausman, W. K. Newey, N. R. Swanson, and T. Woutersen (2014). Testing overidentifying restrictions with many instruments and heteroskedasticity. *Journal of Econometrics 178*, 15–21.

Chetty, R. and N. Hendren (2018). The impacts of neighborhoods on intergenerational mobility II: County-level estimates. *Quarterly Journal of Economics 133*(3), 1163–1228.

Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics 32*(3), 962–994.

El Karoui, N., D. Bean, P. Bickel, C. Lim, and B. Yu (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences 110*(36), 14557–14562.

El Karoui, N. and E. Purdom (2018). Can we trust the bootstrap in high-dimensions? The case of linear models. *Journal of Machine Learning Research 19*(1), 1–66.

Finkelstein, A., M. Gentzkow, and H. Williams (2016). Sources of geographic variation in health care: Evidence from patient migration. *Quarterly Journal of Economics 131*(4), 1681–1726.

Horn, S. D., R. A. Horn, and D. B. Duncan (1975). Estimating heteroscedastic variances in linear models. *Journal of the American Statistical Association 70*(350), 380–385.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Annals of Statistics 1*(5), 799–821.

Jochmans, K. (2020). Heteroscedasticity-robust inference in linear regression models with many covariates. *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2020.1831924.

Kline, P., R. Saggio, and M. Sølvsten (2020). Leave-out estimation of variance components. *Econometrica 88*(5), 1859–1898.

Lachowska, M., A. Mas, R. Saggio, and S. A. Woodbury (2019). Do firm effects drift? Evidence from Washington administrative data. *NBER Working Paper No. 26653*.

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pp. 437–461. Springer.

Phillips, G. D. A. and C. Hale (1977). The bias of instrumental variable estimators of simultaneous equation systems. *International Economic Review 18*(1), 219–228.

Rao, C. R. (1970). Estimation of heteroscedastic variances in linear models. *Journal of American Statistical Association 65*(329), 161–172.

Richard, P. (2019). Residual bootstrap tests in linear models with many regressors. *Journal of Econometrics 208*(2), 367–394.

Romano, J. P., A. M. Shaikh, and M. Wolf (2010). Multiple testing. In *Palgrave Macmillan (eds) The New Palgrave Dictionary of Economics*. Palgrave Macmillan, London.

Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics 116*(2), 681–704.

Sherman, J. and W. J. Morrison (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics 21*(1), 124–127.

Verdier, V. (2020). Estimation and inference for linear models with two-way fixed effects and sparsely matched data. *Review of Economics and Statistics 102*(1), 1–16.

Woodbury, M. A. (1949). The stability of out-input matrices. *Chicago, IL 9*.

Zhou, B., J. Guo, and J.-T. Zhang (2017). High-dimensional general linear hypothesis testing under heteroscedasticity. *Journal of Statistical Planning and Inference 188*, 36–54.