Criterion-Based Inference Without the Information Equality: The Weighted Chi-Square Distribution

Bruce E. Hansen*
University of Wisconsin†

April 2021

Abstract

Criterion-based tests include the F test, the LR test, the GMM distance test, the GMM test for overidentification, the minimum distance test, and the Anderson-Rubin test. Conventional inference with these statistics requires a strong form of correct specification, including the absence of conditional heteroskedasticity, serial correlation, and clustered dependence. More generally, their asymptotic distribution is weighted chi-square, where the weights depend on the eigenvalues of the matrix ratio of the correct asymptotic covariance matrix to the classical (misspecified) covariance matrix. This asymptotic distribution is non-pivotal, but can be consistently estimated, and algorithms are available for its numerical evaluation. We call this implementation the estimated weighted chi-square distribution, and show through a variety of examples that it can be used successfully for accurate asymptotic inference.

*Research support from the NSF and the Phipps Chair are gratefully acknowledged.
†Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706.

1 Introduction

Criterion-based test statistics include the likelihood ratio test, GMM test for overidentification, minimum distance tests, F tests in linear regression, GMM distance (LR-like) tests, and the Anderson-Rubin test. These statistics have several strong advantages. Criterion-based statistics are invariant to parametrization and the algebraic formulation of the null hypothesis. Theoretical and numerical evidence suggests that under correct specification they have superior finite sample performance relative to Wald (delta-method) statistics in nonlinear contexts. For example, Park and Phillips (1988) show that the Edgeworth expansion of the Wald statistic depends on its algebraic formulation, with nonlinearities inducing a nonlinear departure from a leading chi-square approximation, while Hansen (2006) shows that the Edgeworth expansion of the GMM distance statistic is free of the nonlinear terms.

Under correct specification, criterion-based statistics have asymptotic chi-square distributions under the null hypothesis, permitting straightforward inference decisions. However, as we document in Sections 2 and 4, criterion-based tests have the disadvantage that they have non-pivotal asymptotic null distributions under certain forms of misspecification, including heteroskedasticity, clustering, and serial dependence. In contrast, Wald statistics are relatively straightforward to robustify to heteroskedasticity, clustering, and serial dependence. As a consequence, Wald tests dominate applied econometric practice.

As we document in Sections 2 and 4 (and is well-known to specialists), the asymptotic distributions of criterion-based tests in general can be written as weighted sums of chi-square random variables, where the weights are the eigenvalues of (in most examples) a matrix ratio of two covariance matrices. This means that the asymptotic distribution is non-pivotal, but consistently estimable. An asymptotic p-value can be calculated by evaluating the weighted chi-square distribution function at a consistent estimate of the eigenvalues. The resulting p-value is asymptotically $U[0,1]$ distributed as appropriate for accurate inference. Consequently, criterion-based inference using the estimated weighted chi-square distribution is feasible and asymptotically valid.

An alternative to the weighted chi-square distribution is the bootstrap. That is, the bootstrap distribution is in general consistent, permitting asymptotically valid inference. While this is true there are two arguments for considering the weighted chi-square distribution. First, in many cases there are multiple implementations of the bootstrap, leading to potentially contradictory test results. One such issue is how to impose the null hypothesis on a bootstrapped criterion-based test statistic, as there can be multiple methods. Another such issue arises under clustered and time series dependence, where multiple methods are available for bootstrap replication of the serial dependence. In contrast, implementation of the weighted chi-square distribution only requires estimation of asymptotic variance matrices, which is relatively simpler. Second, while the bootstrap distribution is consistent it will not achieve an asymptotic refinement since the asymptotic distribution is non-pivotal. In contrast, the bootstrap applied to our estimated weighted chi-square p-value is a prepivoting statistic as recommended by Beran (1988), which has the potential to achieve an asymptotic refinement. We do not provide a formal theory for such refinements as this would require asymptotic expansions beyond the scope of the present paper. We do, however, include bootstrap methods for comparison in our simulation analysis.

The organization of this paper is as follows. Section 2 describes the structure of the problem and its

solution, using the classical Anderson-Rubin test statistic for illustration. Section 3 discusses numerical calculation of the weighted chi-square cumulative distribution function. Section 4 provides the asymptotic distributions of the following criterion-based statistics: the F statistic, the likelihood ratio statistic, the GMM test for overidentification, the GMM distance statistic, the minimum distance statistic, and the Anderson-Rubin statistic. Section 5 provides a simple simulation experiment for the Anderson-Rubin statistic. Section 6 concludes. Mathematical derivations are presented in the Appendix. Code for computation of the weighted chi-square distribution and all numerical results presented in the paper is posted on the author's webpage.

2 Criterion-Based Inference

Consider the Anderson-Rubin statistic (Anderson and Rubin, 1949) with no included exogenous regressors. The model is $Y_1 = Y_2'\theta + e$ with $\mathbb{E}[e \mid X] = 0$ where X is $q \times 1$. In standard matrix notation the model for a sample of n observations is $\boldsymbol{Y}_1 = \boldsymbol{Y}_2\theta + \boldsymbol{e}$. The Anderson-Rubin statistic, scaled as a Wald statistic, is

$$\mathrm{AR}_n(\theta) = \frac{(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\theta)' \boldsymbol{P} (\boldsymbol{Y}_1 - \boldsymbol{Y}_2\theta)}{(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\theta)' (\boldsymbol{I}_n - \boldsymbol{P}) (\boldsymbol{Y}_1 - \boldsymbol{Y}_2\theta) / (n-q)} \tag{1}$$

where $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$. A test of $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta \neq \theta_0$ rejects for large values of $\mathrm{AR}_n = \mathrm{AR}_n(\theta_0)$. This is a criterion-based statistic because the LIML estimator minimizes $\mathrm{AR}_n(\theta)$.

Under the classical assumption $e \mid X \sim \mathrm{N}(0, \sigma^2)$, AR_n / q has an exact F distribution under $\mathbb{H}_0$. More broadly, if $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2$ then the asymptotic null distribution of AR_n is χ^2_q. However, when the homoskedasticity assumption fails then the asymptotic distribution changes. Indeed, if the observations are i.i.d with finite fourth moments

$$\mathrm{AR}_n \xrightarrow[d]{} \frac{Z' H^{-1} Z}{\sigma^2} = U' V_0^{-1} U$$

where $Z \sim \mathrm{N}(0, \Omega)$, $\Omega = \mathbb{E}\left[XX'e^2\right]$, $H = \mathbb{E}\left[XX'\right]$, $U \sim \mathrm{N}(0, V)$, $V = H^{-1}\Omega H^{-1}$, and $V_0 = \sigma^2 H^{-1}$. We can write $U' V_0^{-1} U$ as $\zeta' B \zeta$ where $\zeta \sim \mathrm{N}(0, I_q)$, $B = C' V_0^{-1} C$, and $V = CC'$ is the Cholesky decomposition of V. Apply the spectral decomposition to the matrix B. This yields $B = P \Lambda P'$ where P is orthonormal and $\Lambda = \mathrm{diag}\{\lambda_1, ..., \lambda_q\}$ contains the eigenvalues of B. Setting $T = P'\zeta \sim \mathrm{N}(0, I_q)$ we find the above expression equals $\sum_{j=1}^{q} \lambda_j Q_j$ where $Q_j = T_j^2$ are independent χ^2_1 random variables. This is a special case of the following distribution.

Definition 1 *For $j = 1, ..., q$, let λ_j be non-negative real numbers, and Q_j be mutually independent χ^2_1 central chi-square random variables. We call*

$$Q(\lambda) = \sum_{j=1}^{q} \lambda_j Q_j$$

*a **weighted chi-square** and write its distribution function as $G(x \mid \lambda)$, where $\lambda = (\lambda_1, ..., \lambda_q)$.*

Our label "weighted chi-square" is a slight simplification of the more common name "weighted sum of chi-squares". More generally, a weighted chi-square can be defined with arbitrary degrees of freedom and non-centrality parameters, but this is not necessary for our development.

We demonstrated above that a quadratic form in normal variables has the distribution $Q(\lambda)$. We state this formally for reference. It is useful to observe that the eigenvalues of the matrix $B = C'V_0^{-1}C$ are equal to those of the matrix $V_0^{-1}CC' = V_0^{-1}V$, which is a less cumbersome expression.

Lemma 1 *If $Q = Z'V_0^{-1}Z$ where $Z \sim \mathrm{N}(0, V)$ then $Q \sim Q(\lambda)$ where $\lambda = (\lambda_1, ..., \lambda_q)$ are the eigenvalues of $V_0^{-1}V$.*

The above derivation shows that the Anderson-Rubin statistic has a weighted chi-square asymptotic null distribution for i.i.d samples with conditionally heteroskedastic errors. It simplifies to a chi-square under homoskedasticity, or a scaled chi-square when $q = 1$, but not otherwise. Our derivation used the assumption of independent observations, but extends to clustered and time series dependence, if the covariance matrix Ω is adjusted to incorporate dependence. The statistic (1) could be robustified to heteroskedasticity by altering the central weight matrix, but as argued by Guggenberger, Kleibergen, and Mavroeidis (2019) this has limited value since this robustified statistic is difficult to extend to allow for subvector inference.

The chi-square weights λ_j which appear in the distribution $Q(\lambda)$ are the eigenvalues of the matrix $V_0^{-1}V$, which is a "matrix ratio" of the correct asymptotic covariance matrix V to the "incorrect" asymptotic covariance matrix V_0 which holds under homoskedasticity. Thus the chi-square weights λ_j measure the discrepancy between the correct covariance matrix and its homoskedastic cousin. This feature of the asymptotic distribution is common to all criterion-based tests, as will be shown in Section 4.

This result implies that under heteroskedastic or dependent errors the Anderson-Rubin test will have incorrect asymptotic size if conventional critical values are used. To see the magnitude of the distortion, take the case $q = 2$ with $\mathbb{E}\left[e^2 \mid X\right] = X_1^2$ where (X_1, X_2) are mutually independent, mean zero, unit variance, $\mathbb{E}\left[X_1^3\right] = 0$, and $\mathbb{E}\left[X_1^4\right] = \mu_4$. Under these assumptions AR_n is asymptotically $Q(\lambda)$ with $\lambda = (1, \mu_4)$. In Table 1 we display the asymptotic Type I error of nominal 5% tests which use the χ^2_2 critical value of 6.0. As we can see, this test has correct asymptotic size when $\mu_4 = 1$, but not for $\mu_4 \neq 1$. When $\mu_4 = 3$ (as when Z_1 is standard normal) the rejection probability is 0.22. When $\mu_4 = 6$ (as when X_1 has a scaled student t distribution with 6 degrees of freedom), then the rejection probability is 0.37. When $\mu_4 = 12$ (the standardized fourth moment of $\log(\mathrm{N}(0, 1/3))$), then the rejection probability is 0.53. These are enormous distortions from the nominal level 0.05. It is worth emphasizing that these are not finite sample distortions but rather are asymptotic.

Table 1: Asymptotic Type I Error of Anderson-Rubin Test

μ_4	1	3	6	12
Asymptotic Size of Nominal 5% Test	0.05	0.22	0.37	0.53

We have used the Anderson-Rubin statistic for illustration due to its simplicity but analogous results apply to all criterion-based statistics. In general, their asymptotic distributions are chi-square under

correct specification and are weighted chi-square under misspecification, heteroskedasticity, clustering, and/or serial dependence. We present details in Section 4.

Correct asymptotic inference can be implemented using the $Q(\lambda)$ distribution if the weights are known, or by replacing the unknown weights with a consistent estimator $\widehat{\lambda}$. For the Anderson-Rubin statistic with independent observations we can set $\widehat{\lambda}$ to equal the eigenvalues[1] of $\widehat{V}_0^{-1}\widehat{V}$ where $\widehat{V}_0 = \widehat{H}^{-1}\widehat{\sigma}^2$ and $\widehat{V} = \widehat{H}^{-1}\widehat{\Omega}\widehat{H}^{-1}$, with

$$\begin{aligned}
\widehat{H} &= n^{-1}\sum_{i=1}^{n} X_i X_i', \\
\widehat{\Omega} &= n^{-1}\sum_{i=1}^{n} X_i X_i' \left(Y_{1i} - \theta_0' Y_{2i}\right)^2, \\
\widehat{\sigma}^2 &= (n-q)^{-1}\sum_{i=1}^{n} \left(Y_{1i} - \theta_0' Y_{2i}\right)^2.
\end{aligned}$$

The estimators $\widehat{\Omega}$ and $\widehat{\sigma}^2$ may alternatively be constructed using a consistent estimator of θ rather than the hypothesized value. Under serial dependence, $\widehat{\Omega}$ should be replaced by a long-run (Newey-West, 1987a) covariance matrix estimator. Under clustering, $\widehat{\Omega}$ should be replaced by a cluster-robust (Arellano, 1987) covariance matrix estimator. Given $\widehat{\lambda} = \left(\widehat{\lambda}_1, ..., \widehat{\lambda}_q\right)$, the estimated asymptotic p-value is $\widehat{p} = 1 - G\left(\mathrm{AR}_n \mid \widehat{\lambda}\right)$ where $G(x \mid \lambda)$ is the weighted chi-square distribution function. Since $\widehat{V}_0^{-1}\widehat{V} \xrightarrow[p]{} V_0^{-1}V$, it follows that $\widehat{p} \xrightarrow[d]{} 1 - G(Q(\lambda) \mid \lambda)$ which has a $U[0,1]$ distribution.

We state a general property for reference, which is a special case of prepivoting as discussed in Beran (1988). The following result follows from the continuous mapping theorem, the fact that $Q(\lambda)$ is continuous in λ, and the probability integral transformation.

Theorem 1 *Suppose that $D_n \xrightarrow[d]{} Q(\lambda)$ where $\lambda = (\lambda_1, ..., \lambda_q)$ are the eigenvalues of a matrix A. Given an estimator $\widehat{A}$ of A, let $\widehat{\lambda} = (\widehat{\lambda}_1, ..., \widehat{\lambda}_q)$ be the eigenvalues of $\widehat{A}$, and set $\widehat{p} = 1 - G\left(D_n \mid \widehat{\lambda}\right)$. If $\widehat{A} \xrightarrow[p]{} A$ then $\widehat{p} \xrightarrow[d]{} U[0,1]$.*

Theorem 1 shows that a criterion-based statistic can be used with heteroskedastic or dependent errors, if the p-value is calculated using the estimated weighted chi-square distribution. This combines the advantages of a criterion-based test with the robustness of a Wald statistic.

3 The Weighted Chi-Square Distribution

In this section we discuss numerical computation of the weighted chi-square cumulative distribution function.

In most cases the best numerical implementation is the series representation due to Ruben (1962) and Farebrother (1984). Define $\lambda_{\min} = \min(\lambda_1, ..., \lambda_q)$, the coefficients $a_m = \sum_{j=1}^{q} \frac{1}{2}\left(1 - \lambda_{\min}/\lambda_j\right)^m$, the

[1] For computation, it is better to first apply the Cholesky decomposition $\widehat{V}_0^{-1} = CC'$ and then calculate the eigenvalues of the symmetric matrix $C'\widehat{V}C$. Equivalently, $\widehat{\lambda}$ can be computed as the generalized eigenvalues of $\widehat{V}$ with respect to $\widehat{V}_0$.

initial condition $b_0 = \prod_{j=1}^{q} \left(\lambda_{\min}/\lambda_j\right)^{1/2}$, and the recursion

$$b_m = \frac{1}{m} \sum_{\ell=0}^{m-1} b_\ell a_{m-\ell}. \tag{2}$$

Ruben's formula for the CDF is

$$G(x \mid \lambda) = \sum_{m=0}^{\infty} b_m G_{q+2m}\left(x/\lambda_{\min}\right). \tag{3}$$

This is a convergent infinite series, written as a weighted sum of chi-square distribution functions. For implementation, the infinite series is truncated with a finite number of terms. A useful feature of (3) is that the accuracy of this truncation can be bounded, so the accuracy of the CDF calculation is controlled.

An implementation of (3) is available in the R function `farebrother` in the package `CompQuadForm`. See Duchesne and Micheaux (2010) for documentation. A Matlab implementation is available on the author's website.

In the special case $q = 2$ the coefficients b_m satisfy the explicit expression

$$b_m = \left(\frac{\lambda_{\min}}{\lambda_{\max}}\right)^{1/2} \left(1 - \frac{\lambda_{\min}}{\lambda_{\max}}\right)^m \frac{\Gamma\left(\frac{1}{2} + m\right)}{\Gamma\left(\frac{1}{2}\right) m!}. \tag{4}$$

The expression (4) can be verified from (2) by induction. Numerical evaluation of (3) with (4) is considerable faster than the recursion (2) and is thus an improvement upon Farebrother's algorithm, but is limited to the case $q = 2$.

While (3) is exact it can be computationally costly in certain extreme situations. These situations arise when the initial condition b_0 is below machine tolerance, or when the number of series terms in (3) needed for convergence is excessive[2]. The latter can occur when $\lambda_{\min}/\lambda_{\max}$ is exceedingly small, or when the intial condition b_0 is extremely small.

In these cases an approximate implementation can be substituted. There is a long literature of computational approximations to $G(x \mid \lambda)$ based on moment matching, including Welch (1938), Satterwaite (1946), Hall (1983), Buckley and Eagleson (1988), and Wood (1989). The most recent and accurate of these approximations is known as LPB4, due to Lindsay, Pilla, and Basak (2000), and is

$$G(x \mid \lambda) \simeq \sum_{m=1}^{p} \pi_m G_k\left(x/\beta_m\right) \tag{5}$$

where $(k, \beta_1, ..., \beta_p, \pi_1, ..., \pi_p)$ are free parameters satisfying $\pi_1 + \cdots + \pi_p = 1$. The parameters are selected to match the first $2p$ moments of $Q(\lambda)$. In principle the accuracy of the approximation can be improved by selecting p large, but there is no specific bound on the approximation error, and the authors recommend $p = 4$ (thus matching the first eight moments). The specific method for determining the parame-

[2]For reference, the calculation (3) can be executed with 5000 series terms in about 0.02 seconds, and with 50,000 series terms in less than 2 seconds.

ters is quite detailed so we do not present this here. For a description, see Bodenham and Adams (2015). An implementation of (5) with $p = 4$ is available in the R function `lpb4` in the package `momentchi2`. A Matlab implementation is available on the author's website.

Bodenham and Adams (2015) provide a thorough examination of the computation speed and accuracy of a variety of the above computational implementations, plus that of Imhof (1961). Bodenham-Adams find that the Farebrother algorithm is the benchmark for computation accuracy. They find that the LPB4 algorithm is somewhat less computationally demanding and somewhat less accurate than the Farebrother algorithm, but is sufficiently accurate for p-value computation. In contrast, they find that for $q < 10$ the Sadderwaithe-Welch, Hall-Buckley-Eagleson, and Wood algorithms are insufficiently accurate for reliable p-value computation. One complication is that they also observed that for $q < 4$ the LPB4 implementation failed to produce a solution in certain parameterizations.

For our numerical implementation, we use a two-step approach. We first attempt the Farebrother (3) series representation[3]. If this fails to converge quickly we switch to the LPB4 algorithm (5). If the LPB4 algorithm fails to produce a solution we re-try the Farebrother algorithm with a larger[4] number of series terms.

4 Examples

The following examples assume a sample of n observations. In each example the parameter is $\theta \in \Theta \subset \mathbb{R}^K$, whose true value θ_0 is presumed to lie in the interior of Θ. The F, likelihood ratio, GMM distance, and minimum distance statistics concern a test of a hypothesis $\mathbb{H}_0 : r(\theta_0) = 0$ against the alternative $\mathbb{H}_1 : r(\theta_0) \neq 0$ given a function $r : \Theta \to \mathbb{R}^q$. The function $r(\theta)$ is assumed continuous in $\theta \in \Theta$, has continuous derivative $R(\theta) = \frac{\partial}{\partial \theta} r(\theta)'$ in a neighborhood of θ_0, and $R = R(\theta_0)$ has full rank q.

For each example we state the asymptotic distribution under high-level conditions. We do this in order to allow for a wide range of conditions, including independent, clustered, and serially dependent observations.

The asymptotic distributions in the following sections are straightfoward to derive by standard methods, but to my knowledge have not been presented in the econometrics literature. We state the results compactly, and present sketches of their proofs in the Appendix.

4.1 F Statistic

Take the model $Y = X'\theta + e$ with $\mathbb{E}[Xe] = 0$ and $\mathbb{E}\left[e^2\right] = \sigma^2 < \infty$. The best linear predictor coefficient is $\theta_0 = \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY]$.

The sum of squared errors function is $S_n(\theta) = \sum_{i=1}^n \left(Y_i - X_i'\theta\right)^2$. The unrestricted and restricted least squares estimators are $\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} S_n(\theta)$ and $\widetilde{\theta} = \underset{\theta \in \Theta : r(\theta) = 0}{\operatorname{argmin}} S_n(\theta)$. The Wald form of the F statistic for $\mathbb{H}_0$

[3] We cap the number of series terms at 100,000 for the case $q = 2$, and at 5000 otherwise.
[4] 100,000.

against $\mathbb{H}_1$ is

$$\mathrm{F}_n = \frac{S_n(\widetilde{\theta}) - S_n(\widehat{\theta})}{\widehat{\sigma}^2}.$$

where $\widehat{\sigma}^2 = (n-K)^{-1}\sum_{i=1}^n \left(Y_i - X_i'\widehat{\theta}\right)^2$.

Under homoskedasticity and no serial or cluster dependence, $\mathrm{F}_n \xrightarrow[d]{} \chi^2_q$. Under broader conditions the asymptotic distribution is a weighted chi-square.

Set $Z_n = n^{-1/2}\sum_{i=1}^n X_i e_i$ and define its asymptotic covariance matrix Ω. Let $\widehat{\Omega}$ be an estimator of Ω. Set $\widehat{H} = n^{-1}\sum_{i=1}^n X_i X_i'$, $\widehat{V}_0 = \widehat{H}^{-1}\widehat{\sigma}^2$, $\widehat{V} = \widehat{H}^{-1}\widehat{\Omega}\widehat{H}^{-1}$, and $\widehat{R} = R(\widehat{\theta})$. Let $\widehat{\lambda} = (\widehat{\lambda}_1, ..., \widehat{\lambda}_q)$ be the eigenvalues of $\left(\widehat{R}'\widehat{V}_0\widehat{R}\right)^{-1}\left(\widehat{R}'\widehat{V}\widehat{R}\right)$ and set $\widehat{p}_{\mathrm{F}} = 1 - G\left(\mathrm{F}_n \mid \widehat{\lambda}\right)$.

Theorem 2 *Assume that (a) $\mathbb{H}_0$ holds; (b) $\widehat{H} \xrightarrow[p]{} H > 0$; (c) $Z_n \xrightarrow[d]{} \mathrm{N}(0,\Omega)$; (d) $\widehat{\sigma}^2 \xrightarrow[p]{} \sigma^2$; and (e) $\widehat{\Omega} \xrightarrow[p]{} \Omega$. Then $\mathrm{F}_n \xrightarrow[d]{} Q(\lambda)$ where $\lambda = (\lambda_1, ..., \lambda_q)$ are the eigenvalues of $\left(R'V_0R\right)^{-1}\left(R'VR\right)$, $V_0 = \sigma^2 H^{-1}$, and $V = H^{-1}\Omega H^{-1}$. Furthermore, $\widehat{p}_{\mathrm{F}} \xrightarrow[d]{} U[0,1]$.*

The matrix V is the asymptotic covariance matrix for $\widehat{\theta}$ under the stated conditions, and allows for heteroskedasticity, clustering, and serial correlation. The matrix V_0 is the classical asymptotic covariance matrix, which holds under conditional homoskedasticity, no clustering, and no serial dependence.

The distribution in Theorem 2 deviates from the chi-square when the correct asymptotic variance V deviates from the "homoskedastic" asymptotic variance V_0. Inference based on the weighted chi-square p-value $\widehat{p}_{\mathrm{F}}$, however, is asymptotically correct, so long as the eigenvalues are calculated from a consistent estimator of the asymptotic covariance matrix.

Thus, to test the hypothesis $\mathbb{H}_0$ against $\mathbb{H}_1$ using an F statistic when the classical regression assumptions are not satisfied, it is asymptotically valid to use the weighted chi-square p-value $\widehat{p}_{\mathrm{F}}$. The hypothesis is rejected at level α if $\widehat{p}_{\mathrm{F}} < \alpha$, and is accepted otherwise.

4.2 Likelihood Ratio Statistic

Take the model $X \sim f(x \mid \theta)$ for some parametric density f. The pseudo-true parameter (regardless of correct specification) is $\theta_0 = \operatorname*{argmin}_{\theta\in\Theta} \mathbb{E}\left[\ell(X,\theta)\right]$.

The negative log-likelihood function is

$$\ell_n(\theta) = \sum_{i=1}^n \ell(X_i,\theta)$$

where $\ell(x,\theta) = -\log f(x \mid \theta)$. The unrestricted and restricted maximum likelihood estimators are $\widehat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \ell_n(\theta)$ and $\widetilde{\theta} = \operatorname*{argmin}_{\theta\in\Theta: r(\theta)=0} \ell_n(\theta)$. The likelihood ratio statistic for $\mathbb{H}_0$ against $\mathbb{H}_1$ is

$$\mathrm{LR}_n = 2\left(\ell_n(\widetilde{\theta}) - \ell_n(\widehat{\theta})\right).$$

Under correct specification, $\mathrm{LR}_n \xrightarrow[d]{} \chi^2_q$. Under broader conditions the asymptotic distribution is a weighted chi-square.

Set $Z_n = n^{-1/2}\sum_{i=1}^n \frac{\partial}{\partial\theta}\ell(X_i,\theta_0)$ and define its asymptotic covariance matrix Ω. Let $\widehat{\Omega}$ be an estimator of Ω. Define $H_n(\theta) = n^{-1}\sum_{i=1}^n \frac{\partial^2}{\partial\theta\partial\theta'}\ell(X_i,\theta)$, $\widehat{H} = H_n(\widehat{\theta})$, $\widehat{V}_0 = \widehat{H}^{-1}$, $\widehat{V} = \widehat{H}^{-1}\widehat{\Omega}\widehat{H}^{-1}$, and $\widehat{R} = R(\widehat{\theta})$. Let $\widehat{\lambda} = (\widehat{\lambda}_1,...,\widehat{\lambda}_q)$ be the eigenvalues of $\left(\widehat{R}'\widehat{V}_0\widehat{R}\right)^{-1}\left(R'\widehat{V}R\right)$ and set $\widehat{p}_{\mathrm{LR}} = 1 - G\left(\mathrm{LR}_n \mid \widehat{\lambda}\right)$.

Theorem 3 *Assume that (a)* $\mathbb{H}_0$ *holds; (b)* $\widehat{\theta} \xrightarrow[p]{} \theta_0$*; (c)* $H_n(\theta) \xrightarrow[p]{} H(\theta)$ *for some* $H(\theta)$ *uniformly in a neighborhood of* θ_0*; (d)* $H = H(\theta_0) > 0$*; (e)* $Z_n \xrightarrow[d]{} \mathrm{N}(0,\Omega)$*; and (f)* $\widehat{\Omega} \xrightarrow[p]{} \Omega$*. Then* $\mathrm{LR}_n \xrightarrow[d]{} Q(\lambda)$ *where* $\lambda = (\lambda_1,...,\lambda_q)$ *are the eigenvalues of* $\left(R'V_0R\right)^{-1}\left(R'VR\right)$*,* $V_0 = H^{-1}$*, and* $V = H^{-1}\Omega H^{-1}$*. Furthermore,* $\widehat{p}_{\mathrm{LR}} \xrightarrow[d]{} U[0,1]$.

The matrix V is the asymptotic covariance matrix for $\widehat{\theta}$ under general conditions, without imposing the information matrix equality $\Omega = H$. The matrix V_0 is the classical asymptotic covariance matrix, and is valid under the information matrix equality. The latter fails under misspecification, clustered dependence, or unmodeled serial dependence.

The distribution $Q(\lambda)$ deviates from the chi-square when the correct asymptotic variance V deviates from the classical asymptotic variance V_0. Inference based on the weighted chi-square p-value $\widehat{p}_{\mathrm{LR}}$, however, is asymptotically correct. Thus, to test the hypothesis $\mathbb{H}_0$ against $\mathbb{H}_1$ using a likelihood ratio statistic when the information matrix equality is not satisfied, use the weighted chi-square p-value $\widehat{p}_{\mathrm{LR}}$. The hypothesis is rejected at level α if $\widehat{p}_{\mathrm{LR}} < \alpha$ and is accepted otherwise.

4.3 GMM Overidentification Test

Take the moment model $\mathbb{E}\left[g(X,\theta_0)\right] = 0$ for some $\ell \times 1$ function $g(x,\theta)$ where $\ell \geq k$. The GMM criterion is

$$J_n(\theta) = n\overline{g}_n(\theta)'W_n\overline{g}_n(\theta)$$
$$\overline{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^n g(X_i,\theta)$$

where $W_n > 0$ is an $\ell\times\ell$ weight matrix. The unrestricted and restricted GMM estimators are $\widehat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}}\, J_n(\theta)$ and $\widetilde{\theta} = \underset{\theta\in\Theta:r(\theta)=0}{\operatorname{argmin}}\, J_n(\theta)$. When $\ell > k$ the GMM statistic for overidentifying restrictions (proposed by Lars Hansen (1982)) is $J_n = J_n(\widehat{\theta})$.

Set $Z_n = n^{-1/2}\sum_{i=1}^n g(X_i,\theta_0)$ and define its asymptotic covariance matrix Ω. Let $\widehat{\Omega}$ be an estimator of Ω. Define $G_n(\theta) = n^{-1}\sum_{i=1}^n \frac{\partial}{\partial\theta'}g(X_i,\theta)$, $\widehat{G} = G_n(\widehat{\theta})$, and

$$\widehat{W}_G = W_n - W_n\widehat{G}\left(\widehat{G}'W_n\widehat{G}\right)^{-1}\widehat{G}'W_n.$$

Let $\widehat{\lambda} = (\widehat{\lambda}_1,...,\widehat{\lambda}_{\ell-K})$ be the non-zero eigenvalues of $\widehat{\Omega}\widehat{W}_G$ and set $\widehat{p}_J = 1 - G\left(J_n \mid \widehat{\lambda}\right)$.

Theorem 4 *Assume that (a)* $\widehat{\theta} \xrightarrow[p]{} \theta_0$*; (b)* $G_n(\theta) \xrightarrow[p]{} G(\theta)$ *for some* $G(\theta)$ *uniformly in a neighborhood of* θ_0*; (c)* $G = G(\theta_0)$ *has full rank* k*; (d)* $Z_n \xrightarrow[d]{} \mathrm{N}(0,\Omega)$*; (e)* $W_n \xrightarrow[p]{} W > 0$*; and (f)* $\widehat{\Omega} \xrightarrow[p]{} \Omega$*. Then* $J_n \xrightarrow[d]{} Q(\lambda)$

where $\lambda = (\lambda_1, ..., \lambda_{\ell-K})$ *are the non-zero eigenvalues of* ΩW_G *where*

$$W_G = W - WG\left(G'WG\right)^{-1} G'W.$$

Furthermore, $\widehat{p}_J \xrightarrow[d]{} U[0,1]$.

When W equals the efficient weight matrix Ω^{-1} then $\Omega W_G = I_\ell - G\left(G'\Omega^{-1}G\right)^{-1} G'\Omega^{-1}$ which has unit eigenvalues and $Q(\lambda)$ simplifies to $\chi^2_{\ell-K}$. Otherwise the distribution is a weighted chi-square.

The matrices $\widehat{W}_G$ and W_G have rank $\ell - K$ so the matrices $\widehat{\Omega}\widehat{W}_G$ and ΩW_G have $\ell - K$ non-zero eigenvalues.

The weighted chi-square distribution in Theorem 4 for the GMM overidentification test takes a different form from the hypothesis tests, as the eigenvalues λ are not based on the matrix ratios of two covariance matrices.

Theorem 4 shows that when GMM estimation has been performed using a GMM criterion where the weight matrix W_n is not necessarily asymptotically efficient (for example, does not take into account clustering), then it is appropriate to use the weighted chi-square p-value $\widehat{p}_J$. The overidentifying restrictions are rejected at level α if $\widehat{p}_J < \alpha$ and are accepted otherwise.

4.4 GMM Distance Test

The GMM distance statistic (proposed by Newey and West (1987b)) for $\mathbb{H}_0$ against $\mathbb{H}_1$ is $D_n = J_n(\widetilde{\theta}) - J_n(\widehat{\theta})$. If W_n is consistent for the efficient weight matrix, then $D_n \xrightarrow[d]{} \chi^2_q$. Otherwise the asymptotic distribution is a weighted chi-square. Inference based on the weighted chi-square p-value $\widehat{p}_J$, however, is asymptotically correct.

Define $\widehat{V}_0 = \left(\widehat{G}'W_n\widehat{G}\right)^{-1}$,

$$\widehat{V} = \left(\widehat{G}'W_n\widehat{G}\right)^{-1}\left(\widehat{G}'W_n\widehat{\Omega}W_n\widehat{G}\right)\left(\widehat{G}'W_n\widehat{G}\right)^{-1},$$

and $\widehat{R} = R(\widehat{\theta})$. Let $\widehat{\lambda} = (\widehat{\lambda}_1, ..., \widehat{\lambda}_q)$ be the eigenvalues of $\left(\widehat{R}'\widehat{V}_0\widehat{R}\right)^{-1}\left(\widehat{R}'\widehat{V}\widehat{R}\right)$ and set $\widehat{p}_D = 1 - G\left(D_n \mid \widehat{\lambda}\right)$.

Theorem 5 *Assume that the conditions of Theorem 4 hold, plus* $\mathbb{H}_0$. *Then* $D_n \xrightarrow[d]{} Q(\lambda)$ *where* $\lambda = (\lambda_1, ..., \lambda_q)$ *are the eigenvalues of* $\left(R'V_0R\right)^{-1}\left(R'VR\right)$, *with* $V_0 = \left(G'WG\right)^{-1}$ *and*

$$V = \left(G'WG\right)^{-1}\left(G'W\Omega WG\right)\left(G'WG\right)^{-1}.$$

Furthermore, $\widehat{p}_D \xrightarrow[d]{} U[0,1]$.

The matrix V is the asymptotic covariance matrix for $\widehat{\theta}$, and the matrix V_0 is the asymptotic covariance matrix when W is the efficient weight matrix Ω^{-1}. The distribution $Q(\lambda)$ deviates from the chi-square when the weight matrix W deviates from the efficient weight matrix. Inference based on the weighted chi-square p-value $\widehat{p}_D$, however, is asymptotically correct.

Thus, when estimation is performed by GMM when the weight matrix W_n is not necessary asymptotically efficient, then the GMM distance statistic can be still used for inference if assessed using the estimated weighted chi-square p-value $\widehat{p}_D$. The hypothesis is rejected at level α if $\widehat{p}_D < \alpha$ and accepted otherwise.

4.5 Minimum Distance

Let $\widehat{\theta}$ be a first-stage estimator with estimator $\widehat{V}$ of its asymptotic covariance matrix V. Let $W_n > 0$ be a weight matrix. The minimum distance criterion is

$$\mathrm{MD}_n(\theta) = n\left(\widehat{\theta} - \theta\right)' W_n \left(\widehat{\theta} - \theta\right).$$

The minimum distance estimator under $\mathbb{H}_0$ is $\widetilde{\theta} = \underset{\theta \in \Theta: r(\theta)=0}{\operatorname{argmin}} J_n(\theta)$. The minimum distance statistic for $\mathbb{H}_0$ against $\mathbb{H}_1$ is $\mathrm{MD}_n = \mathrm{MD}_n(\widetilde{\theta})$.

Set $\widehat{R} = R(\widehat{\theta})$. Let $\widehat{\lambda} = (\widehat{\lambda}_1, ..., \widehat{\lambda}_q)$ be the eigenvalues of $\left(\widehat{R}' W_n \widehat{R}\right)^{-1}\left(\widehat{R}' \widehat{V} \widehat{R}\right)$ and set $\widehat{p}_{\mathrm{MD}} = 1 - G\left(\mathrm{MD}_n \mid \widehat{\lambda}\right)$.

Theorem 6 *Assume that (a) $\mathbb{H}_0$ holds; (b) $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow[d]{} \mathrm{N}(0, V)$; (c) $W_n \xrightarrow[p]{} W > 0$; and (d) $\widehat{V} \xrightarrow[p]{} V$. Then $\mathrm{MD}_n \xrightarrow[d]{} Q(\lambda)$ where $\lambda = (\lambda_1, ..., \lambda_q)$ are the eigenvalues of $\left(R' W R\right)^{-1}\left(R' V R\right)$.*

When the weight matrix is set as $W_n = \widehat{V}^{-1}$ then the asymptotic distribution of MD_n is chi-square. If the weight matrix is set otherwise then the asymptotic distribution of MD_n is weighted chi-square. Therefore if estimation is performed by minimum distance with a weight matrix not set to equal the inverse of the covariance matrix, then inference can be based on the estimated weighted chi-square p-value $\widehat{p}_{\mathrm{MD}}$. The hypothesis is rejected at level α if $\widehat{p}_{\mathrm{MD}} < \alpha$ and accepted otherwise.

4.6 Anderson-Rubin

In Section 2 we used the Anderson-Rubin statistic in a model with no exogenous regressors to illustrate criterion-based inference with the weighted chi-square distribution. In this section we present the empirically-relevant case with exogenous regressors.

The model is $Y_1 = Y_2'\theta + X_1'\gamma + e$ with $\mathbb{E}[e \mid X] = 0$ for $X = [X_1, X_2]$, where the included regressors X_1 are $K \times 1$ and the excluded instruments X_2 are $q \times 1$. The Anderson-Rubin statistic, scaled as a Wald statistic, is

$$\mathrm{AR}_n(\theta) = \frac{(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\theta)' \boldsymbol{P}_U (\boldsymbol{Y}_1 - \boldsymbol{Y}_2\theta)}{(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\theta)' (\boldsymbol{I}_n - \boldsymbol{P}) (\boldsymbol{Y}_1 - \boldsymbol{Y}_2\theta) / (n - K - q)}$$

where $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$, $\boldsymbol{P}_U = \boldsymbol{U}\left(\boldsymbol{U}'\boldsymbol{U}\right)^{-1}\boldsymbol{U}'$ and $\boldsymbol{U} = \boldsymbol{X}_2 - \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2$. A test of $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta \neq \theta_0$ rejects for large values of $\mathrm{AR}_n = \mathrm{AR}_n(\theta_0)$. This is a criterion-based statistic because the LIML estimator minimizes $\mathrm{AR}_n(\theta)$.

If the error is normally distributed then AR_n / q has an exact F distribution under $\mathbb{H}_0$. More broadly, under homoskedasticity and no serial dependennce then the asymptotic null distribution of AR_n is χ^2_q.

The Anderson-Rubin statistic is typically motivated as a method for confidence set construction. A confidence set for θ equals the set of parameter values such that $\mathrm{AR}_n(\theta)$ is less than a critical value. If conventional critical values are used this is only valid under homoskedasticity and serial independence. If weighted chi-square critical values are used then such confidence sets have broader validity.

Set $Z_n = n^{-1/2}\sum_{i=1}^n X_i e_i$ and define its asymptotic covariance matrix Ω. Let $\widehat{\Omega}$ be an estimator of Ω. Let $\widehat{\sigma}^2$ be the residual variance estimator from the regression of $Y_1 - Y_2'\theta$ on X. Set $\widehat{H} = n^{-1}\sum_{i=1}^n X_i X_i'$, $\widehat{V}_0 = \widehat{H}^{-1}\widehat{\sigma}^2$, and $\widehat{V} = \widehat{H}^{-1}\widehat{\Omega}\widehat{H}^{-1}$. Set $S = (0_{q\times K}, I_q)'$. Let $\widehat{\lambda} = (\widehat{\lambda}_1, ..., \widehat{\lambda}_q)$ be the eigenvalues of $\left(S'\widehat{V}_0 S\right)^{-1}\left(S'\widehat{V}S\right)$ and set $\widehat{p}_{\mathrm{AR}} = 1 - G\left(\mathrm{AR}_n \mid \widehat{\lambda}\right)$.

Theorem 7 *Assume that (a)* $\widehat{H} \xrightarrow[p]{} H > 0$*; (b)* $Z_n \xrightarrow[d]{} \mathrm{N}(0,\Omega)$*; (c)* $\widehat{\sigma}^2 \xrightarrow[p]{} \sigma^2$*; and (d)* $\widehat{\Omega} \xrightarrow[p]{} \Omega$*. Then* $\mathrm{AR}_n \xrightarrow[d]{} Q(\lambda)$ *where* $\lambda = (\lambda_1, ..., \lambda_q)$ *are the eigenvalues of* $\left(S'V_0S\right)^{-1}\left(S'VS\right)$*,* $V_0 = \sigma^2 H^{-1}$*, and* $V = H^{-1}\Omega H^{-1}$*. Furthermore,* $\widehat{p}_{\mathrm{AR}} \xrightarrow[d]{} U[0,1]$.

Theorem 7 shows that it is asymptotically valid to use the weighted chi-square p-value $\widehat{p}_{\mathrm{AR}}$ for inference using the Anderson-Rubin statistic, allowing for violation of the classical assumptions on the errors. This allows application to a broad set of contexts, including heteroskedastic errors, serially dependent errors, and clustered dependence.

As mentioned above, the typical motivation for the Anderson-Rubin statistic is confidence region construction. This can be done using the estimated weighted chi-square distribution, allowing for heteroskedasticity, serial correlation, and/or clustering. Given the estimated eigenvalues $\widehat{\lambda}$, an ξ level confidence region for θ is the set of values such that $G\left(\mathrm{AR}_n(\theta) \mid \widehat{\lambda}\right) \leq \xi$. The eigenvalues $\widehat{\lambda}$ may be calculated based on an estimator $\widehat{\theta}$ of θ, or could be calculated $\widehat{\lambda}(\theta)$ separately for each θ.

5 Simulation

We illustrate the use of the weighted chi-square distribution in a simple simulation experiment. Following Section 2 we consider the Anderson-Rubin (1949) statistic in a linear model with heteroskedastic errors. We compare inference based on the F, weighted chi-square, and three bootstrap distributions.

The model is $Y_1 = Y_2'\theta + e$ with $\mathbb{E}[e \mid X] = 0$ where X is $q \times 1$. There are no included exogenous regressors. The Anderson-Rubin statistic for a test of $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta \neq \theta_0$ is (1). We generate the model error as $e \mid X \sim \mathrm{N}(0, X_1^2)$ to induce heteroskedasticity, vary q among $(2,5,8)$, and vary sample size n among $(100,500)$. The instruments $X_2, ..., X_q$ are generated as independent $\mathrm{N}(0,1)$.

The instrument X_1 is generated from one of four designs, constructed to create increasingly asymmetric eigenvalues λ. The designs are:

1. $X_1 = 1$. This is identical to specifying e as conditionally homoskedastic.
2. $X_1 \sim \mathrm{N}(0,1)$.
3. $X_1 \sim t(6)$.
4. $X_1 \sim \exp\left(\mathrm{N}(0, 1/3)\right)$.

In designs 3 & 4 the instrument X_1 was re-normalized to have zero mean and unit variance.

30,000 simulated replications were made for each parameterization, the Anderson-Rubin statistic (1) calculated, and p-values calculated by five methods: F distribution, estimated weighted chi-square distribution, and three bootstrap methods. We calculated the finite sample size of nominal 0.05 tests and report the results in Table 2.

The first section of Table 2 shows the size of the Anderson-Rubin statistic with critical values taken from the F distribution. We see that the test is over-sized except in Design 1 which has homoskedastic errors. In Designs 2, 3, and 4, the size distortion is large, with the Type I error reaching as high as 48%. The distortion worses as the sample size increases.

Table 2: Size of Nominal 5% Tests

	$n = 100$			$n = 500$		
	$q = 2$	$q = 5$	$q = 8$	$q = 2$	$q = 5$	$q = 8$
F						
Design 1	0.050	0.051	0.049	0.049	0.051	0.051
Design 2	0.209	0.161	0.134	0.212	0.162	0.142
Design 3	0.306	0.241	0.209	0.335	0.279	0.245
Design 4	0.411	0.338	0.304	0.482	0.404	0.372
Weighted Chi-Square						
Design 1	0.046	0.039	0.032	0.048	0.049	0.046
Design 2	0.038	0.029	0.020	0.047	0.042	0.040
Design 3	0.035	0.025	0.017	0.043	0.041	0.037
Design 4	0.024	0.018	0.012	0.038	0.033	0.033
Regression Bootstrap						
Design 1	0.050	0.049	0.043	0.049	0.052	0.051
Design 2	0.055	0.051	0.042	0.052	0.050	0.048
Design 3	0.068	0.060	0.053	0.058	0.057	0.055
Design 4	0.087	0.079	0.068	0.070	0.063	0.064
Hall-Horowitz Bootstrap						
Design 1	0.042	0.026	0.011	0.048	0.046	0.043
Design 2	0.029	0.015	0.005	0.047	0.040	0.036
Design 3	0.022	0.011	0.004	0.042	0.039	0.034
Design 4	0.014	0.007	0.002	0.036	0.031	0.028
Prepivot Bootstrap						
Design 1	0.038	0.045	0.042	0.049	0.051	0.051
Design 2	0.014	0.032	0.026	0.047	0.046	0.045
Design 3	0.029	0.026	0.021	0.039	0.039	0.037
Design 4	0.019	0.017	0.014	0.028	0.026	0.026

Note: Rejection frequencies from 30,000 simulation draws.

Note: The asymptotic standard error for any entry of 0.05 is 0.001.

The second section of Table 2 shows the size of the Anderson-Rubin test using the estimated weighted

chi-square distribution and its p-value $\widehat{p}_{\text{AR}}$. To estimate Ω we use the efficient estimator

$$\widehat{\Omega} = n^{-1}\sum_{1}^{n} X_i X_i' \left(Y_{1i} - Y_{2i}'\theta_0\right)^2.$$

We can see that in all entries the Type I error does not exceed the nominal level of 5%. There is, however, conservative size distortion, with actual Type I error rates ranging between 1% and 5%. This size distortion is increasing in the number of instruments q, with the severity of the design, and is decreasing with the sample size n. The source of the finite sample size distortion appears to be the estimation of the weights λ. Weight estimation can result in spurious estimated weight heterogeneity, resulting in spuriously conservative p-values. This distortion decreases with larger samples where weight estimation is more accurate.

An alternative to the estimated weighted chi-square distribution is the bootstrap, but there are multiple implementations of the bootstrap. We focus on the standard nonparametric bootstrap which resamples i.i.d. from the observations. For bootstrap testing a key issue is the need to impose the null hypothesis on the bootstrap statistic. For our first bootstrap implementation we observe that the Anderson-Rubin statistic can be written as a regression F test. Specifically, (1) equals

$$\text{AR}_n = \frac{\widehat{\beta}'\left(\boldsymbol{X}'\boldsymbol{X}\right)\widehat{\beta}}{\widehat{\sigma}^2}$$

where $\widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\theta_0\right)\right)$ is the least squares coefficient from the regression of $Y_1 - Y_2'\theta_0$ on X, and $\widehat{\sigma}^2$ is the associated residual variance estimator. This is a regression F test of the hypothesis $\mathbb{H}_0 : \beta = 0$ in the regression model $Y_1 - Y_2'\theta_0 = X'\beta + e$. The standard bootstrap version of this regression F statistic is

$$\text{AR}_n^* = \frac{\left(\widehat{\beta}^* - \widehat{\beta}\right)'\left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*\right)\left(\widehat{\beta}^* - \widehat{\beta}\right)}{\widehat{\sigma}^{*2}}$$

where the starred statistics are calculated on the bootstrap sample. This bootstrap statistic imposes the null hypothesis $\beta = 0$ by centering the regression estimator $\widehat{\beta}^*$ at the sample value $\widehat{\beta}$. We calculate 1000 bootstrap simulated statistics AR_n^* for each simulation replication, and calculated the bootstrap p-value as the percentage of the bootstrap AR_n^* which exceed the value AR_n. The bootstrap test rejects at the 5% level if the bootstrap p-value is smaller than 0.05.

The third section of Table 2 shows the size of this bootstrap test, labeled as "Regression Bootstrap". The size of the test is much improved relative to the F distribution, but still has considerable size distortion, with actual Type I error rates ranging between 4.2% and 8.7%.

An alternative method to impose the null hypothesis was proposed by Hall and Horowitz (1996) in the context of GMM estimation. They proposed the statistic

$$\text{AR}_n^{**} = \frac{\left(\boldsymbol{X}^{*\prime}\left(\boldsymbol{Y}_1^* - \boldsymbol{Y}_2^*\theta_0\right) - \boldsymbol{X}'\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\theta_0\right)\right)\left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*\right)^{-1}\left(\boldsymbol{X}^{*\prime}\left(\boldsymbol{Y}_1^* - \boldsymbol{Y}_2^*\theta_0\right) - \boldsymbol{X}'\left(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\theta_0\right)\right)}{\widehat{\sigma}^{*2}}.$$

This bootstrap statistic imposes the null hypothesis $\mathbb{E}\left[Xe\right] = 0$ by centering the bootstrap moment $\boldsymbol{X}^{*\prime}\left(\boldsymbol{Y}_1^* - \boldsymbol{Y}_2^*\theta_0\right)$

at the sample value $\boldsymbol{X}'(\boldsymbol{Y}_1 - \boldsymbol{Y}_2\theta_0)$. Otherwise the implementation of bootstrap inference is the same. There is no theoretical reason to expect AR_n^* or AR_n^{**} to have different finite sample performance.

The fourth section of Table 2 shows the size of this bootstrap test, labeled as "Hall-Horowitz Bootstrap". Surprisingly, the size of the test is considerably different from the regression bootstrap test. The Hall-Horowitz test is highly conservative, especially in the smaller sample, for larger q, and for the stronger designs. The actual Type I error rates ranging between 0.0% and 4.8%. This highly conservative performance means that this test will have low power relative to alternative statistics.

The fact that the size of the two bootstrap tests based on AR_n^* and AR_n^{**} are noticably different from one another is disconcerting, pointing to a meaningful divergence between bootstrap implementations. This is not a caution against use of the bootstrap; rather, it is pointing out that bootstrap methods are not a panacea for finite sample inference.

Our final bootstrap implementation is Beran (1988) prepivoting applied to the estimated weighted chi-square p-value $\widehat{p}_{\mathrm{AR}}$. On each bootstrap sample we calculated the regression bootstrap statistic AR_n^*, the covariance matrix ratio $\widehat{V}_0^{*-1}\widehat{V}^*$ using the same formula as on the original observations, its eigenvalues $\widehat{\lambda}^*$, and the bootstrap p-value statistic $\widehat{p}_{\mathrm{AR}}^* = 1 - G\left(\mathrm{AR}_n^* \mid \widehat{\lambda}^*\right)$. The prepivot bootstrap p-value is the percentage of the bootstrap statistics $\widehat{p}_{\mathrm{AR}}^*$ which are smaller than the sample value $\widehat{p}_{\mathrm{AR}}$. The bootstrap test rejects at the 5% level if the bootstrap p-value is smaller than 5%. The size of this test is displayed in the fifth section of Table 2. The results are similar to those for the non-bootstrapped weighted chi-square distribution. The rejection rates are more conservative for $q = 2$, but the reverse for most entries for $q = 5$ and $q = 8$.

Comparing the five tests reported in Table 2 we can make the following conclusions. First, under heteroskedasticity the Anderson-Rubin statistic with classical F critical values exhibits large size distortions. Second, the over-rejection rates can be effectively eliminated by use of the estimated weighted chi-square distribution. Third, the latter exhibits conservative size distortions in small samples. Fourth, while the regression bootstrap dramatically reduces the finite sample distortions, they are not eliminated, and the performance of the bootstrap does not dominate the performance of the estimated weighted chi-square distribution. Overall the estimated weighted chi-square p-values have the best performance yet are computationally inexpensive.

6 Conclusion

Criterion-based tests have many advantages relative to Wald tests, but have the traditional disadvantage that their asymptotic distributions are difficult to robustify to misspecification, including heteroskedasticity, serial correlation, and clustering. In such contexts the asymptotic distribution of criterion-based statistics is weighted chi-square. As we show, these weights are consistently estimable, and combined with the Farebrother algorithm provides an asymptotically valid method for p-value calculation and inference. This pairing – criterion-based tests with the estimated chi-square distribution function – combines the best advantages of criterion-based and Wald-based tests.

7 Appendix: Mathematical Proofs

Proof of Theorem 2: The unconstrained estimator $\widehat{\theta}$ is consistent for θ_0. As the set $\{\theta : r(\theta) = 0\}$ is compact and $r(\theta_0) = 0$ by assumption, it follows that the constrained estimator $\widetilde{\theta}$ is also consistent for θ_0. The restricted estimator satisfies $r\left(\widetilde{\theta}\right) = 0$. Expanding each element in a first order Taylor expansion about θ_0, we find

$$0 = r\left(\widetilde{\theta}\right) = r(\theta_0) + R_n'\left(\widetilde{\theta} - \theta_0\right) = R_n'\left(\widetilde{\theta} - \theta_0\right) \tag{6}$$

where each row of R_n equals the corresponding row of $R(\theta^*)$ for some θ^* on the line segment joining $\widetilde{\theta}$ and θ_0. We deduce from (6) that

$$R_n'\widetilde{\theta} = R_n'\theta_0. \tag{7}$$

Furthermore, since $\widetilde{\theta}$ is consistent then so is θ^*, and thus $R_n \xrightarrow[p]{} R$ by the continuous mapping theorem.

The constrained estimator $\widetilde{\theta}$ minimizes the Lagrangian

$$\frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\theta)'(\boldsymbol{Y} - \boldsymbol{X}\theta) + \lambda' r(\theta)$$

which has first order condition

$$0 = -\boldsymbol{X}'\left(\boldsymbol{Y} - \boldsymbol{X}\widetilde{\theta}\right) + \widetilde{R}\widetilde{\lambda}$$

where $\widetilde{R} = R\left(\widetilde{\theta}\right)$. Premultiplying by $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ we obtain

$$0 = -\widehat{\theta} + \widetilde{\theta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\widetilde{R}\widetilde{\lambda} \tag{8}$$

and premultiplying by R_n' and using (7) we obtain

$$0 = -R_n'\left(\widehat{\theta} - \theta_0\right) + R_n'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\widetilde{R}\widetilde{\lambda}.$$

Solving for $\widetilde{\lambda}$ and inserting into (8) we obtain

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta} - \widetilde{\theta}\right) &= \sqrt{n}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\widetilde{R}\left(R_n'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\widetilde{R}\right)^{-1}R_n'\left(\widehat{\theta} - \theta_0\right) \\ &= \widehat{H}^{-1}\widetilde{R}\left(R_n'\widehat{H}^{-1}\widetilde{R}\right)^{-1}R_n'\widehat{H}^{-1}Z_n.\end{aligned}$$

By algebraic manipulations, we find that

$$\begin{aligned}S_n(\widetilde{\theta}) - S_n(\widehat{\theta}) &= \left(\widetilde{\theta} - \widehat{\theta}\right)'\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\widetilde{\theta} - \widehat{\theta}\right) \\ &= Z_n'\widehat{H}^{-1}R_n\left(R_n'\widehat{H}^{-1}\widetilde{R}\right)^{-1}\widetilde{R}'\widehat{H}^{-1}\widetilde{R}\left(R_n'\widehat{H}^{-1}\widetilde{R}\right)^{-1}R_n'\widehat{H}^{-1}Z_n.\end{aligned} \tag{9}$$

The assumptions imply that

$$\sqrt{n}R_n'\widehat{H}^{-1}Z_n \xrightarrow[d]{} Z \sim \mathrm{N}(0, R'VR).$$

Applied to (9) we obtain

$$S_n(\widetilde{\theta}) - S_n(\widehat{\theta}) \xrightarrow[d]{} Z'\left(R'H^{-1}R\right)^{-1}Z.$$

We also see that $\widehat{\sigma}^2 \underset{p}{\longrightarrow} \sigma^2$. Together,

$$\mathrm{F}_n = \frac{S_n(\widetilde{\theta}) - S_n(\widehat{\theta})}{\widehat{\sigma}^2} \underset{d}{\longrightarrow} Z' \left(R' V_0 R\right)^{-1} Z.$$

By Lemma 1 this is distributed $Q(\lambda)$ where $\lambda = (\lambda_1, ..., \lambda_q)$ are the eigenvalues of $\left(R' V_0 R\right)^{-1} \left(R' V R\right)$.

The assumptions imply that $\left(\widehat{R}' \widehat{V}_0 \widehat{R}\right)^{-1} \left(\widehat{R}' \widehat{V} \widehat{R}\right) \underset{p}{\longrightarrow} \left(R' V_0 R\right)^{-1} \left(R' V R\right)$. Thus by Theorem 1, $\widehat{p}_{\mathrm{F}} \underset{d}{\longrightarrow} U[0,1]$ as stated. ■

Proof of Theorem 3: By the consistency of the estimators, a second order Taylor series expansion of $\ell_n(\widetilde{\theta})$ about $\widehat{\theta}$, the fact $\frac{\partial}{\partial \theta} \ell_n(\widehat{\theta}) = 0$, and the assumptions on $H_n(\theta)$, we can show that

$$\mathrm{LR}_n = 2\left(\ell_n(\widetilde{\theta}) - \ell_n(\widehat{\theta})\right) = \sqrt{n}\left(\widetilde{\theta} - \widehat{\theta}\right)' H \sqrt{n}\left(\widetilde{\theta} - \widehat{\theta}\right) + o_p(1). \tag{10}$$

By a Lagrange multiplier analysis similar to that of the proof of Theorem 2, it is straightforward to show that

$$\sqrt{n}\left(\widetilde{\theta} - \widehat{\theta}\right) = -H^{-1} R \left(R' H^{-1} R\right)^{-1} R' \sqrt{n}\left(\widehat{\theta} - \theta_0\right) + o_p(1).$$

By a standard Taylor series argument and the assumptions,

$$R' \sqrt{n}\left(\widehat{\theta} - \theta_0\right) = -R' H^{-1} Z_n + o_p(1) \underset{d}{\longrightarrow} Z \sim \mathrm{N}\left(0, R' V R\right).$$

These three equations and $H^{-1} = V_0$ combine to show that

$$\mathrm{LR}_n = \sqrt{n}\left(\widehat{\theta} - \theta_0\right)' R \left(R' H^{-1} R\right)^{-1} R' \sqrt{n}\left(\widehat{\theta} - \theta_0\right) + o_p(1) \underset{d}{\longrightarrow} Z' \left(R' V_0 R\right)^{-1} Z.$$

By Lemma 1 this is distributed $Q(\lambda)$ where $\lambda = (\lambda_1, ..., \lambda_q)$ are the eigenvalues of $\left(R' V_0 R\right)^{-1} \left(R' V R\right)$.

The assumptions imply that $\left(\widehat{R}' \widehat{V}_0 \widehat{R}\right)^{-1} \left(\widehat{R}' \widehat{V} \widehat{R}\right) \underset{p}{\longrightarrow} \left(R' V_0 R\right)^{-1} \left(R' V R\right)$. Thus by Theorem 1, $\widehat{p}_{\mathrm{LR}} \underset{d}{\longrightarrow} U[0,1]$ as stated. ■

Proof of Theorem 4: By a standard Taylor series expansion of the first-order-condition for $\widehat{\theta}$ about θ_0 and standard manipulations we find that

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) = -\left(G' W G\right)^{-1} G' W Z_n + o_p(1) \underset{d}{\longrightarrow} -\left(G' W G\right)^{-1} G' W Z \tag{11}$$

where $Z \sim \mathrm{N}(0, \Omega)$. By a Taylor expansion of $\overline{g}_n(\widehat{\theta})$ about θ_0 and the above result we find

$$\begin{aligned} \sqrt{n}\overline{g}_n(\widehat{\theta}) &= Z_n + G' \sqrt{n}\left(\widehat{\theta} - \theta_0\right) + o_p(1) \\ &= \left(I_\ell - G' \left(G' W G\right)^{-1} G' W\right) Z_n + o_p(1) \\ &\underset{d}{\longrightarrow} \left(I_\ell - G' \left(G' W G\right)^{-1} G' W\right) Z. \end{aligned} \tag{12}$$

Hence

$$J_n = \sqrt{n}\overline{g}_n(\widehat{\theta})' W_n \sqrt{n}\overline{g}_n(\widehat{\theta}) \xrightarrow[d]{} Z'\left(W - WG'\left(G'WG\right)^{-1}G'W\right)Z.$$

By Lemma 1 this is distributed $Q(\lambda)$ where $\lambda = (\lambda_1, ..., \lambda_{\ell-K})$ are the eigenvalues of

$$\Omega\left(W - WG'\left(G'WG\right)^{-1}G'W\right).$$

The assumptions imply that $\widehat{A} \xrightarrow[p]{} A$. Thus by Theorem 1, $\widehat{p}_J \xrightarrow[d]{} U[0,1]$ as stated. ■

Proof of Theorem 5: By a Taylor series expansion of $\overline{g}_n(\widetilde{\theta})$ about $\widehat{\theta}$ we obtain

$$\sqrt{n}\overline{g}_n(\widetilde{\theta}) = \sqrt{n}\overline{g}_n(\widehat{\theta}) + G\sqrt{n}\left(\widetilde{\theta} - \widehat{\theta}\right) + o_p(1).$$

Thus

$$\begin{aligned}
D_n &= n\overline{g}_n(\widetilde{\theta})' W_n \overline{g}_n(\widetilde{\theta}) - n\overline{g}_n(\widehat{\theta})' W_n \overline{g}_n(\widehat{\theta}) \\
&= n\left(\widetilde{\theta} - \widehat{\theta}\right)' GWG\left(\widetilde{\theta} - \widehat{\theta}\right) + 2n\left(\widetilde{\theta} - \widehat{\theta}\right) G'W\overline{g}_n(\widehat{\theta}) + o_p(1) \\
&= n\left(\widetilde{\theta} - \widehat{\theta}\right)' GWG\left(\widetilde{\theta} - \widehat{\theta}\right) + 2n\left(\widetilde{\theta} - \widehat{\theta}\right) G'W\left(I_\ell - G'\left(G'WG\right)^{-1}G'W\right)Z_n + o_p(1) \\
&= n\left(\widetilde{\theta} - \widehat{\theta}\right)' GWG\left(\widetilde{\theta} - \widehat{\theta}\right) + o_p(1).
\end{aligned}$$

The third equality is (12), and the fourth uses the fact $G'W\left(I_\ell - G'\left(G'WG\right)^{-1}G'W\right) = 0$.

By a Lagrange multiplier analysis similar to that of the proof of Theorem 2 it is straightforward to show that

$$\sqrt{n}\left(\widehat{\theta} - \widetilde{\theta}\right) = (GWG)^{-1}R\left(R'(GWG)^{-1}R\right)^{-1}R'\sqrt{n}\left(\widehat{\theta} - \theta_0\right) + o_p(1).$$

(11) implies $R'\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow[d]{} Z \sim \mathrm{N}\left(0, R'VR\right)$. Together, we find

$$\begin{aligned}
D_n &= \sqrt{n}\left(\widehat{\theta} - \theta_0\right)R\left(R'(GWG)^{-1}R\right)^{-1}R'\sqrt{n}\left(\widehat{\theta} - \theta_0\right) + o_p(1) \\
&\xrightarrow[d]{} Z'\left(R'V_0R\right)^{-1}Z
\end{aligned}$$

By Lemma 1 this is distributed $Q(\lambda)$ where $\lambda = (\lambda_1, ..., \lambda_q)$ are the eigenvalues of $\left(R'V_0R\right)^{-1}\left(R'VR\right)$.

The assumptions imply that $\left(\widehat{R}'\widehat{V}_0\widehat{R}\right)^{-1}\left(\widehat{R}'\widehat{V}\widehat{R}\right) \xrightarrow[p]{} \left(R'V_0R\right)^{-1}\left(R'VR\right)$. Thus by Theorem 1, $\widehat{p}_D \xrightarrow[d]{} U[0,1]$ as stated. ■

Proof of Theorem 6: By a Lagrange multiplier analysis similar to that of the proof of Theorem 2 it is straightforward to show that

$$\sqrt{n}\left(\widehat{\theta} - \widetilde{\theta}\right) = W^{-1}R\left(R'WR\right)^{-1}R'\sqrt{n}\left(\widehat{\theta} - \theta_0\right) + o_p(1).$$

The assumptions imply that $R'\sqrt{n}\left(\widehat{\theta}-\theta_0\right) \underset{d}{\longrightarrow} Z \sim \mathrm{N}\left(0, R'VR\right)$. Together we find

$$\begin{aligned} J_n &= n\left(\widehat{\theta}-\widetilde{\theta}\right)' W_n \left(\widehat{\theta}-\widetilde{\theta}\right) + o_p(1) \\ &= \sqrt{n}\left(\widehat{\theta}-\theta_0\right)' R\left(R'WR\right)^{-1} R'\sqrt{n}\left(\widehat{\theta}-\theta_0\right) + o_p(1) \\ &\underset{d}{\longrightarrow} Z'\left(R'WR\right)^{-1} Z. \end{aligned}$$

By Lemma 1 this is distributed $Q(\lambda)$ where $\lambda = (\lambda_1, ..., \lambda_q)$ are the eigenvalues of $\left(R'WR\right)^{-1}\left(R'VR\right)$.

The assumptions imply that $\left(\widehat{R}'W_n\widehat{R}\right)^{-1}\left(\widehat{R}'\widehat{V}\widehat{R}\right) \underset{p}{\longrightarrow} \left(R'WR\right)^{-1}\left(R'VR\right)$. Thus by Theorem 1, $\widehat{p}_{\mathrm{MD}} \underset{d}{\longrightarrow} U[0,1]$ as stated. ■

Proof of Theorem 7: The statistic $\mathrm{AR}_n(\theta_0)$ equals the F statistic for $\beta_2 = 0$ in the linear regression $Y_1 - Y_2'\theta_0 = X_1'\beta_1 + X_2'\beta_2 + e$ with θ_0 known. This can be written as

$$\mathrm{AR}_n = Z_n'\widehat{H}^{-1}S\left(S'\widehat{V}_0S\right)^{-1}S'\widehat{H}^{-1}Z_n.$$

The assumptions imply

$$S'\widehat{H}^{-1}Z_n \underset{d}{\longrightarrow} Z \sim \mathrm{N}\left(0, S'VS\right)$$

where $Z \sim \mathrm{N}\left(0, S'VS\right)$. Thus

$$\mathrm{AR}_n \underset{d}{\longrightarrow} Z'\left(S'V_0S\right)^{-1}Z.$$

By Lemma 1 this is distributed $Q(\lambda)$ where $\lambda = (\lambda_1, ..., \lambda_q)$ are the eigenvalues of $\left(S'V_0S\right)^{-1}\left(S'VSR\right)$.

The assumptions imply that $\left(S'\widehat{V}_0S\right)^{-1}\left(S'\widehat{V}S\right) \underset{p}{\longrightarrow} \left(S'V_0S\right)^{-1}\left(S'VS\right)$. Thus by Theorem 1, $\widehat{p}_{\mathrm{AR}} \underset{d}{\longrightarrow} U[0,1]$ as stated. ■

References

[1] Anderson, Theodore W. and Herman Rubin (1949): "Estimation of the parameters of a single equation in a complete system of stochastic equations," *The Annals of Mathematical Statistics*, 20, 46-63.

[2] Arellano, Manuel (1987): "Computing robust standard errors for within-groups estimators," *Oxford Bulletin of Economics and Statistics*, 49, 431-434.

[3] Beran, Rudolf (1988): "Prepivoting test statistics: A bootstrap view of asymptotic refinements," *Journal of the American Statistical Association*, 83, 687-697.

[4] Bodenham, Dean A. and Niall M. Adams (2015): "A comparison of efficient approximations for a weighted sum of chi-squared random variables," *Statistics and Computing*, 26, 917-928.

[5] Buckley, M. J., and G. K. Eagleson (1988): "An approximation to the distribution of quadratic forms in normal random variables," *Australian Journal of Statistics*, 30, 150-159.

[6] Duchesne, Pierre and Pierre Lafaye 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.

[7] Farebrother, R. W. (1984): "Algorithm AS 204: The distribution of a positive linear combination of χ^2 random variables," *Journal of the Royal Statistical Society, Series C*, 33, 332-339.

[8] Guggenberger, Patrik, Frank Kleibergen, and Sophocles Mavroeidis (2019): "A more powerful subvector Anderson Rubin test in linear instrumental variables regression," *Quantitative Economics*, 10, 487-526.

[9] Hall, Peter (1983): "Chi squared approximations to the distribution of a sum of independent random variables," *Annals of Probability*, 11, 1028-1036.

[10] Hansen, Bruce E. (2006): "Edgeworth expansions for the Wald and GMM statistics for nonlinear restrictions," *Econometric Theory and Practice*, edited by Dean Corbae, Steven N. Durlauf, and Bruce E. Hansen, Cambridge University Press.

[11] Hansen, Lars P. (1982): "Large sample properties of generalized method of moments estimators." *Econometrica*, 50, 1029-1054.

[12] Hall, Peter, and Joel L. Horowitz (1996): "Bootstrap critical values for tests based on generalized-method-of-moments estimators," *Econometrica*, 64, 891-916.

[13] Imhof, J. P. (1961): "Computing the distribution of quadratic forms in normal variables," *Biometrika*, 48, 419-426.

[14] Lindsay, Bruce G., Ramani S. Pilla, and Prasanta Basak (2000): "Moment-based approximations of distributions using mixtures: Theory and applications," *Annals of the Institute of Statistical Mathematics*, 52, 215-230.

[15] Newey, Whitney K. and Kenneth D. West (1987a): "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance marix," *Econometrica*, 55, 703-708.

[16] Newey, Whitney K. and Kenneth D. West (1987b): "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, 28, 777-787.

[17] Park, Joon Y. and Peter C. B. Phillips (1988): "On the formulation of Wald tests of nonlinear restrictions," *Econometrica*, 56, 1065-1083.

[18] Ruben, Harold (1962): "Probability content of regions under spherical normal distributions, IV: The distribution of homogeneous and non-homogeneous quadratic functions of normal variables," *Annals of Mathematical Statistics*, 33, 542-570.

[19] Satterthwaite, F. E. (1946): "An approximate distribution of estimates of variance components," *Biometrics Bulletin*, 2, 110-114.

[20] Welch, B. L. (1938): "The significance of the difference between two means when the population variances are unequal," *Biometrika*, 29, 350-362.

[21] Wood, Andrew T. A. (1989): "An F approximation to the distribution of a linear combination of chi-squared variables," *Communications in Statistics – Simulation and Computation*, 18, 1439-1456.