Practical Steps to Improve Specification Testing

Practical Steps to Improve Specification Testing*

Austin Nichols

Abt Associates

Mark E. Schaffer

Heriot-Watt University

September 2020

Abstract

Testing of null hypotheses about model parameters, with a dichotomous choice of reject or not, has come under heavy fire from many quarters of late, but the same attention has not been paid to specification tests. We outline a framework for improving specification tests that accepts the dichotomous choice of rejecting a model or not in a very small class of cases (where results are entirely uninterpretable upon rejecting the null, as in the case of underidentification), but argues for more informative testing in most cases. The Arellano (1993) version of Hausman's (1978) specification test is one example where results can be more informative. We also provide a novel example of more informative testing in the case of White's test of heteroskedasticity. In general, we argue that the researcher should generically not reject a model based on misspecification tests, but should describe the sensitivity of estimation or inference to potential violations of assumptions using more informative specification tests.

*Invited paper for the Fourth International Econometric Conference of Vietnam, 'Prediction and Causality in Econometrics and Related Topics', Banking University of Ho-Chi-Minh City, Vietnam, 11-13 January 2021. All errors are our own.

1 Introduction

Misspecification tests (or equivalently, specification tests) are a mainstay of applied econometrics, in both practice and pedagogy. A researcher wishing to publish results using, say, the random effects panel estimator, or a linear instrumental variable estimator, could not do so without mention of the misspecification tests of the maintained assumptions in those models.[1] Any econometrics textbook, undergraduate or advanced, that did not cover misspecification tests and how to interpret them, would be criticized as incomplete and inadequate. Specification tests in econometrics are reported in terms of a test of a null hypothesis that the specified model is correct, where rejection of the null hypothesis is interpreted to mean that one or more of the model assumptions fail, and the model should be rejected as incorrect. We argue the dichotomous misspecification test is not a useful formulation, and provide constructive examples of an alternative approach that focuses on the utility of the model even in situations where the model is not perfect.

The use of "null hypothesis significance testing" (NHST) and p-values has attracted an enormous amount of attention in the applied statistics literature in recent years, most of it critical (and most of the rest attempts to defend it from this criticism). For example, the American Statistical Association released a "Statement on Statistical Significance and P-Values" in 2016 (Wasserstein and Lazar, 2016), with six principles including "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold." In 2019 *Nature* published a paper by Amrhein et al. (2019), cosigned by over 800 researchers (including one of us), suggesting that researchers "retire statistical significance" in favor of more nuanced interpretation. This debate has focused on null hypothesis testing of model parameters, and for good reasons: this is what researchers are almost always interested in. In these settings, some form of interval estimation is almost always preferred to a dichotomous "reject or not" decision about a null hypothesis.

Curiously, this critical literature seems mostly to have ignored misspecification testing. This is surprising, since many of the criticisms of dichotomous "reject or not" testing clearly apply equally or even more strongly to misspecification test. A researcher should not blithely assume that if the p-value of the fixed-vs-random effects test is 0.06, all is well with the world and the random effects estimator can be used

[1] Sargan (1975), in a passage popularized by Godfrey (1988) in his book on misspecification tests, famously characterized a particular use of OLS without testing assumptions as a "pious fraud" (p. 321). Sargan's comments are worth citing at length. He begins his short piece with "... Our usual problem in econometric work is to detect when we are troubled with mis-specification, and to have the imagination to envisage what types of mis-specification may be leading our estimates astray. We do not know how we should specify our model, and we shall be misled by our estimates unless we specify it correctly." (p. 321). He concludes by saying, "... there are a range of tests for mis-specification (including those for serial correlation) which are not very often used in in applied work. Despite the problems associated with 'data-mining' I consider that a suggested specification should be tested in all possible ways, and only those specifications which survive and correspond to a reasonable economic model should be used" (p. 322). Pagan (1990) in reviewing Godfrey's book says "we use conventions all the time and we can be very dependent upon their validity... we would want to see if there is evidence against these assumptions in our data, and it is for this reason that mis-specification testing has become an important part of research" (p. 273). Similarly, Wooldridge (1990) begins with the sentence "Specification testing has become an integral part of the econometric model building process" (p. 17).

without hesitation, nor should they assume that if the p-value is 0.04 the random effects estimator is useless and uninformative. If we reject homogeneity of variance, or an intracluster correlation coefficient of zero, with a p-value of 0.04, we not should not assume our confidence intervals are uninformative—nor if we fail to reject should we assume the models are perfect. But how should a researcher respond in these cases to an observed test statistic? Should the researcher discard the entire research project, or switch to another model (and in this case, report both, or only the preferred model)? Or should they report differently from the model that failed the test? We argue that the answer to this question varies with the misspecification test at hand, and provide some constructive advice for specific cases.

In this paper we look at misspecification testing from the perspective of the 'retire statistical significance' critics. We look at several examples: Ramsey's RESET test, Hausman's misspecification test, White's test for heteroskedasticity, and tests of instrument relevance in IV/GMM estimation. Our main recommendation is that misspecification tests should, if possible, be cast in a metric that is immediately useful and informative to the researcher. The first choice of metric is that of the parameter of interest: if a model assumption fails, can this failure be expressed in terms of the size of the bias in the parameter of interest? This will not always be possible: an example is a test for heteroskedasticity, where failure of the homoskedasticity assumption has implications for the variance estimator, so inference is affected but not point estimates. We show how to recast White's famous heteroskedasticity test (a staple of applied econometrics in both teaching and practice) in terms of the size of the bias in the variance estimator. Finally, we also show that some misspecification tests in common use need no modification, because they are already in a metric that is readily interpreted or because the implications for the researcher's decision are unambiguous. That is, the 'retire statistical significance' critique applies only selectively to misspecification testing.

2 Misspecification testing

A misspecification test is a test of some or all of the assumptions of the model used by the researcher. Under the null hypothesis that the model is well-specified, the test statistic has a known distribution. The usual way a misspecification test is used is that the researcher performs the test, calculates the p-value, compares this to some preselected significance level, and on that basis either rejects or fails to the reject the null hypothesis. We call this 'Null Hypothesis Misspecification Testing' or NHMT, by analogy to NHST.

Exactly what is being rejected is partly up to preferred interpretation of the researcher. The researcher may have doubts about a particular assumption of the model and may choose to interpret the rejection of the null as the rejection of that particular assumption. The test may also be designed to have power to

detect violations of a specific assumption or set of assumptions on which the model relies. An alternative is for the researcher to treat the test as an "omnibus" test and treat a rejection as a rejection of the entire model. This last approach is common, as it is not always clear what set of assumptions is included in the null. As Wooldridge (1990) points out, tests are often "characterized by the same feature: validity of the tests requires imposition of more than just the hypotheses of interest under H_0" (p.18).

The primary problem with this approach, as already mentioned, is that the dichotomous procedure is fundamentally arbitrary. The selected p-value, say five percent, is typically chosen via rule of thumb, and is not tied to integrated a loss function over a posterior estimate.[2] The use of an arbitrary p-value for the decision to reject or not to reject is unconnected to the consequences of model specification failure. It is perfectly possible, for example, for the researcher to reject the model even though the implied bias in the estimated coefficient is economically minuscule.

A second, more subtle problem is that model assumptions are often not believed by the researcher to be literally true. For example, the model may be assumed to be linear, but linearity is merely believed to be an acceptable approximation. Or the disturbance is assumed to be normal but normality is just an acceptable approximation. The researcher might assume there is no unobserved time-invariant heterogeneity in the panel but zero heterogeneity is just an acceptable approximation, or that regressors are assumed to be orthogonal to the disturbance terms, but exact orthogonality is just an acceptable approximation. So the researcher already knows the null hypothesis cannot be true, but wants to know if maintaining the hypothesis as true as an assumption is harmful to estimation or inference, and NHMT does not answer this question—it answers the question to which the research already knows the answer.

It follows that in these examples, if the researcher has enough data, the researcher will detect violations of the null and "reject the model" simply because they have enough information in the dataset to identify that these are just approximations.

In homage to Arthur Goldberger (who famously said that multicollinearity should rather be called "micronumerosity" because it was a problem resulting from not enough data), we could call this problem "macronumerosity", i.e. the researcher has 'too much data', and rejects the model because they are detecting that acceptable approximations are indeed just approximations. Even with a tiny implied bias or size distortion, the test with near-infinite data would almost surely reject, and tell us to discard the model and learn nothing from the data.

In these circumstances, we suggest that researchers should instead, if possible, frame the misspecification test in the metric of a parameter of interest. For example, an interval estimate for the size of the possible bias in the estimated parameter is straightforward for the researcher to interpret. This may not always be possible, however. It may sometimes be the case that no such interpretation is available, and

[2] [https://www.youtube.com/watch?v=XX1IWVVpZ7A]

practical use of the test is hard to recommend. On the other hand, sometimes a straightforward reject/fail to reject decision following the misspecification test may actually be easy to interpret.

We now examine in detail several commonly-used misspecification tests from this perspective.

2.1 Ramsey's RESET test

Ramsey's (1969) RESET or "regression specification-error test" is commonly found in econometric textbooks, probably in part because the intuition is straightforward and easy to explain to students. It is sometimes described as a test for omitted variables, but is better described as a test for omitted nonlinearities or interactions.

The researcher begins with the standard linear model,

$$y_i = X_i\beta + \varepsilon_i \tag{1}$$

but may suspect that the true model is actually

$$y_i = X_i\beta + W_i\alpha + \varepsilon_i \tag{2}$$

where W_i here consists of higher-order elements and interactions of X_i. A null hypothesis misspecification test for these nonlinearities is simply $H_0 : \alpha = 0$, implemented by estimating the augmented equation by OLS and testing $\hat{\alpha}$ using a Wald test.

The test can rapidly lose power as the order and number of interactions increases, and so a common approach is to estimate the original equation using OLS and obtain the fitted values $\hat{y}_i$. The equation is then augmented by higher-order terms of $\hat{y}_i$ and then estimated by OLS, e.g.,

$$y_i = X_i\beta + \gamma_2\hat{y}_i^2 + \gamma_3\hat{y}_i^3 + \gamma_4\hat{y}_i^4 + \eta_i \tag{3}$$

and the NHST for nonlinearities is $H_0 : \gamma_2 = \gamma_3 = \gamma_4 = 0$.

The test is simple to conduct, simple to understand, and simple to teach. It is also a good illustration of a misspecification test that actually tells the researcher very little. In particular, the test falls foul of the macronumerosity problem. With real-world data, it is very unlikely that the "true model" is exactly linear. With enough data, a researcher will be able (eventually) to reject the null. But this does not mean that the original linear model was misspecified—it may have been a perfectly acceptable approximation. Even if there is an "important" neglected nonlinearity or interaction, the test does not tell us what form it might take, whereas a principled kernel-based semiparametric estimator or LASSO might indicate a

simpler parametric model that applies.[3]

A second important point this example raises is the test needs to address the researcher's objective, whether it is predictive or causal inference. If the objective is prediction, then the test amounts to a procedure that can detect unmodeled nonlinearities or interactions. But there are better, more modern methods for improving predictive performance (penalized regression, cross-validation, etc.) and the test tells us nothing about which nonlinearities or interactions are missing. If the objective is causal inference, then rejection (or not) of the null doesn't actually have any necessary implications for whether the parameter of interest β is consistently estimated or not. The simplest way to illustrate this is with a well-known example, namely the linear probability model (LPM). The LPM is a simple and robust method for estimating the conditional mean (see e.g. Angrist and Pischke (2009) for a spirited defence). The RESET misspecification test will, with enough data, be able to detect the fact that a linear model is being used to estimate a nonlinear DGP. But the fact that the linear model is not technically correct in this situation does not diminish its usefulness.

We note, by the way, that despite its ease of use, RESET is also only rarely used in econometric practice. We suspect its use is rare because the null is so easy to reject the null in practice, and the implication of the rejection so useless in practice, and not because of perceived flaws in the rationale behind the test.

2.2 Testing for heteroskedasticity

Testing for heteroskedasticity is a staple of econometrics textbooks at both the undergraduate and graduate level. White's (1980) general test is often the basis of the presentation. There are good pedagogical reasons for this. In particular, White's test is closely related to the construction of heteroskedastic-consistent estimators of $var(\hat{\beta})$, and the relationship between White's general test and the robust covariance estimator illuminates both. That said, the standard use of the test, with a null hypothesis that conditional heteroskedasticity is not present, tells the researcher little that is actually useful. Later we will revisit this test with a proposal to make it useful. Here we outline the test and why, as typically used, it is uninformative. Our exposition follows that in Hayashi (2000).

The basic linear model with the elements needed to explicate the test are as follows:

$$y_i = x_i\beta + \varepsilon_i \tag{4}$$

$$\Sigma_{xx} = E(x_i x_i') \tag{5}$$

[3] See e.g. Cattaneo and Jansson (2018).

$$S = E(\varepsilon_i^2 x_i x_i') \tag{6}$$

$$Avar(\widehat{\beta})_{OLS} = \Sigma_{xx}^{-1} S \Sigma_{xx}^{-1} \tag{7}$$

The asymptotic variance of $\widehat{\beta}_{OLS}$ is consistently estimated by

$$\widehat{Avar(\widehat{\beta}_{OLS})} = S_{xx}^{-1} \widehat{S} S_{xx}^{-1} \tag{8}$$

where

$$S_{xx} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i' \tag{9}$$

and $\widehat{S}$ is some consistent estimate of S.

The classical "observed information matrix" (OIM) and heteroskedastic-consistent (HC) estimators of S are

$$\widehat{S}_{OIM} = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2\right)\left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i'\right) = \hat{\sigma}^2 S_{xx} \qquad \widehat{S}_{HC} = \frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 x_i x_i' \tag{10}$$

where $\hat{\varepsilon}$ is the residual from some consistent estimator of β.

The intuition behind White's general test, and what makes it so useful for pedagogical purposes, follows directly from the observation that $\widehat{S}_{OIM}$ requires the assumption of homoskedasticity to be consistent for S, and $\widehat{S}_{HC}$ does not. If the difference between $\widehat{S}_{OIM}$ and $\widehat{S}_{HC}$ is "large", it is an indication - assuming, of course, that other assumptions are not violated - that ε is conditionally heteroskedastic.

To construct the general heteroskedasticity test, define Ψ_i to be the vector with the m unique and nonconstant elements of $x_i x_i'$. Under conditional homoskedasticity, the difference between $\widehat{S}_{OIM}$ and $\widehat{S}_{HC}$ goes to zero:

$$\widehat{S}_{HC} - \widehat{S}_{OIM} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\varepsilon}_i^2 - \hat{\sigma}^2) x_i x_i' \underset{p}{\rightarrow} \mathbf{0}_{K \times K} \tag{11}$$

Define c_n to be an $m \times 1$ vector:

$$c_n \equiv \frac{1}{n} \sum_{i=1}^{n} (\hat{\varepsilon}_i^2 - \hat{\sigma}^2) \Psi_i \underset{p}{\rightarrow} \mathbf{0}_{m \times 1} \tag{12}$$

Under the null that ε is conditionally homoskedastic, $\sqrt{n} c_n$ converges in probability to a normal variate with mean zero and asymptotic variance B. Given some $\widehat{B}$ consistent for B, White's test for heteroskedasticity is:

$$n\, c_n' \widehat{B}^{-1} c_n \underset{d}{\to} \chi^2(m) \tag{13}$$

White suggests two estimators of B:

$$\widehat{B} = \frac{1}{n}\sum_{i=1}^{n} [\hat{\varepsilon}_i^2 - \hat{\sigma}^2]^2 (\Psi_i - \bar{\Psi})'(\Psi_i - \bar{\Psi}) \tag{14}$$

$$\widehat{B}_{nR2} = \frac{1}{n}\sum_{i=1}^{n} [\hat{\varepsilon}_{in}^2 - \hat{\sigma}_n^2]^2 \;\; \frac{1}{n}\sum_{i=1}^{n} \Big((\Psi_i - \bar{\Psi})'(\Psi_i - \bar{\Psi}) \Big) \tag{15}$$

where $\bar{\Psi}$ is the sample mean of Ψ_i. The variance estimator $\widehat{B}_{nR2}$ assumes homokurtosis and is the familiar version obtained as an nR^2 statistic from an artificial regression of the squared residuals on the elements of Ψ.

Variations on the test include the selection of elements of Ψ. The Breusch-Pagan/Godfrey test, for example, uses only levels of x_i in the artificial regression using the squared residuals. Another variation is to replace the elements of Ψ_i with the fitted values of $\hat{y}_i$, i.e., a specific linear combination of the elements of x_i. Yet another variation is to use $\hat{y}_i$ and $\hat{y}_i^2$ as the indicator variables in the artificial regression.

There are several issues here:

- As with the RESET test, White's general test and its variants, as typically taught and used in practice, falls foul of the macronumerosity problem. With real-world data it is very unlike that the true variance S of the score vector is perfectly homoskedastic. With enough data, a researcher will be able (eventually) to reject the null of homoskedasticity. But this does not mean that the use of the classical covariance estimators $\widehat{S}_{OIM}$ and $\widehat{Avar(\widehat{\beta}_{OLS})}_{OIM} = S_{xx}^{-1}\widehat{S}_{OIM}S_{xx}^{-1} = \hat{\sigma}^2 S_{xx}^{-1}$ is inappropriate. The size distortions could easily be extremely small.

- Indeed, the nice properties of $\widehat{S}_{HC}$ are asymptotic, and it is quite possible that the size distortions from finite sample bias in $\widehat{S}_{HC}$ could lead to rejection of the null.

- The test quickly loses power as K increases, because it "looks in every direction" (the degrees of freedom increase with K^2). This is a well-known problem.

- The test, as typically presented, uses all the elements of x_i. White's general test uses all the levels, squares and cross-products; the Breusch-Pagan/Godfrey version uses all the levels; the $\hat{y}_i$ variant combines all the levels into a single indicator. But the researcher typically is interested in performing inference on just one, or perhaps a few, elements of the parameter vector β. Many of the regressors are often control variables where inference is uninteresting or unneeded. Yet if differences between the elements of $\widehat{S}_{OIM}$ and $\widehat{S}_{HC}$ corresponding to some control variables are "large", the test will still reject ... even if the differences between elements corresponding to the regressors of interest are

small.

- c_n is a difference vector in the same metric as the variance of the score S, not the variance of the parameter vector $Avar(\hat{\beta})$. The researcher is, or should be, concerned about size distortions when performing inference using the latter, but the test is capturing large differences in the former. This makes the test harder to interpret.

2.3 The Hausman test

Hausman's (1978) misspecification test is possibly the best-known misspecification test in econometrics. Standard practice in both teaching and applications is to present the test in terms of NHMT ("null hypothesis misspecification testing"). This is uninformative and immediately falls foul of the macronumerosity problem. But the nature of the Hausman test is such that addressing this problem is very straightforward, because the test is already in the most useful metric possible: that of the parameters themselves.

We illustrate with the standard textbook example of "fixed vs. random effects". Write the basic error components model as

$$y_{it} = x_{it}\beta + \alpha_i \eta_i t \tag{16}$$

where α_i is some time-invariant panel-specific effect. The fixed effects estimator exploits only the orthogonality conditions $E(x_{it}\eta_i t) = 0$. The random effects estimator is more efficient because in addition it also exploits the orthogonality conditions $E(x_{it}\alpha_i) = 0$. In most applications, the researcher is more confident that the 'within' orthogonality conditions $E(x_{it}\eta_i t) = 0$ hold compared to the 'between' orthogonality conditions $E(x_{it}\alpha_i) = 0$. The Hausman test applied to fixed vs. random effects is usually interpreted as a test whether the between orthogonality conditions hold.

The classical Hausman test formulation is a vector-of-contrasts test:

$$H = n\left(\widetilde{\beta}_{FE} - \widehat{\beta}_{RE}\right)' \left(\mathbf{V}(\widetilde{\beta}_{FE}) - \mathbf{V}(\widehat{\beta}_{RE})\right)^{-1} \left(\widetilde{\beta}_{FE} - \widehat{\beta}_{RE}\right) \tag{17}$$

where $\mathbf{V}(\widetilde{\beta}_{FE})$ and $\mathbf{V}(\widehat{\beta}_{RE})$ are estimators of the asymptotic variances of the inefficient fixed effects estimator $\widetilde{\beta}_{FE}$ and efficient random effects estimator $\widehat{\beta}_{RE}$, respectively. Under $H_0 : E(x_{it}\alpha_i) = 0$ and the maintained assumption that $E(x_{it}\eta_i t) = 0$, the Hausman statistic H is distributed as $\chi^2(K)$ where K is the dimension of β.

The problem here is macronumerosity: with real world data, the "between" orthogonality conditions $H_0 : E(x_{it}\alpha_i) = 0$ are unlikely to be exactly true. With a large enough sample size and real data, the researcher will reject H, even if the vector of contrasts between $\widetilde{\beta}_{FE}$ and $\widetilde{\beta}_{RE}$ is extremely small in practical terms.

It makes much more sense to interpret the Hausman test in the metric of the test itself (the parameter estimates) than to use the uninformative NHMT formulation. This is in fact easy to do, and follows naturally from the extension of the test to the non-i.i.d. case due to Arellano (1993). This version of the test is implemented via an artificial regression, and the regression coefficients provide a natural interpretation of the test results.

Take our basic error-components model and include the panel time means[4] $\overline{\mathbf{x}}_i'$ as regressors:

$$y_{it} = \mathbf{x}_{it}'\beta + \overline{\mathbf{x}}_i'\gamma + \alpha_i\eta_i t \tag{18}$$

Estimate this equation using the RE estimator. The $\widehat{\beta}$ produced by RE estimation of this equation is **exactly** the same as the $\widehat{\beta}_{FE}$, the estimate obtained by FE estimation of the original equation (without the Mundlak fixed effects). Moreover, a Wald test of $H_0 : \gamma = 0$ using the RE estimation is **exactly** the same as the Hausman H (when the FE estimate of $\widehat{\sigma}_\eta^2$ is used). This is because the estimates of Mundlak coefficients γ are just the differences between the fixed and random effects of β.

The key point for our purposes is that this artificial regression comes with standard errors and confidence intervals for γ. The research can therefore immediately assess just how important in practical terms are any differences between $\widetilde{\beta}_{FE}$ and $\widetilde{\beta}_{RE}$. The test is easily extended to the non-i.i.d. case by using an appropriate covariance estimator (e.g., the cluster-robust covariance estimator).

2.4 Testing for instrument relevance: Examples of good practice in NHMT

Most NHST (null hypothesis significance testing) in science falls foul of the critiques discussed above, and it is hard to think of examples where the test of a null of the form $H_0 : \beta = \beta_0$ is an interesting answer to an scientific question, as we usually want to know the range of plausible magnitudes for a parameter, not whether the probability of observing our data given one hypothesized value and a host of other assumptions falls below some line. The situation is not quite so bleak with NHMT (null hypothesis misspecification testing), and there are examples where standard practice is also good practice. In this section we consider one such category, namely testing for instrument relevance.

A key assumption required for consistency by IV/GMM estimators is the rank condition:

$$rank(E(z_i x_i)) = K$$

where K is number of regressors (the column dimension of x_i). If $E(z_i x_i)$ is less than full column rank, the IV/GMM estimator is underidentified and hence inconsistent for β.

[4]The $\overline{\mathbf{x}}_i$ are sometimes called "Mundlak fixed effects"; see Mundlak (1978).

Classical tests for the rank of a matrix go back to Anderson's (1951) canonical correlations test. Extensions to the non-i.i.d. were introduced starting in the 1990s, notably by Cragg and Donald (1993) and Kleibergen and Paap (2006). Windmeijer (2018) extends the set of underidentification tests still further and shows their relationship to tests of overidentification.

The null hypothesis for a rank test of underidentification is $H_0 : rank(E(z_i x_i)) = K - 1$, i.e., a rank reduction of 1. What makes this test useful is the clarity of failing to reject the null. If the null of underidentification cannot be rejected at the chosen significance level, the implication for the researcher is that model cannot be used for any inference. Evidently, it's time to go back to the drawing board, and in this situation, one should not report any results (contra the fourth principle of the ASA Statement on Statistical Significance and P-Values "Proper inference requires full reporting and transparency", in which "null" results need to be published to ensure that the distribution of published estimates is not artificially truncated, to avoid bias in research synthesis).

Rejecting the null of underidentification is not enough for the researcher to conclude that all is well, however; the model may still be weakly identified. A literature beginning with Staiger and Stock (1997) has looked at the implications for estimation by IV/GMM when the the correlation between the endogenous regressors and the excluded instruments are weak. Interestingly, discarding estimates where instruments are judged to be weak introducing pernicious bias in the research synthesis, as shown by Andrews et al. (2019).

Stock and Yogo (2005) introduced two such tests for the i.i.d. case. These tests use the i.i.d. version of the underidentification test statistic of Cragg and Donald (1993), which is essentially a Wald version of Anderson's (1951) test; see Windmeijer (2018). However, the weak identification tests use different critical values for this statistic (based on its simulated distribution) that depend on the estimator used (IV or LIML) and, most importantly for our purposes, whether the researcher is testing for bias or size distortion. It is the fact that the null hypotheses of these tests are in interpretable metrics that makes them examples of an appropriate use of NHMT.

To execute the Stock-Yogo maximal bias test, the researcher chooses a critical value according to the number of endogenous regressors K, the number of instruments L, the maximal bias of the IV estimator relative to OLS, and the desired significance level.

Stock and Yogo (2005) provide a simple illustration: for the case of a single endogenous regressor $K = 1$, a maximal bias of 10%, and a 5% significance level, for all L the critical value is approximately 11. A Cragg-Donald test statistic that exceeds 11 implies rejecting the null that the worst-case bias is 10% (i.e., that it is less than 10%). Stock and Yogo point out that this provides a formal justification for the Staiger-Stock 'rule of thumb' that the Cragg-Donald statistic should be 10 or more. Note that the null hypothesis is in a metric that is natural and important for the researcher, namely the bias of the IV estimator

that results from weak identification.

The Stock-Yogo maximal size test sets critical values according to the size distortion that the researcher is willing to accept. For example, when conducting inference on the estimated coefficients, the researcher may choose the usual 5% significance level. To conduct the Stock-Yogo maximal size test, the researcher would specify that they are willing to tolerate a maximal size of 15%, i.e., a distortion of up to 10% for a 5% level test.

With the full distribution in hand of the critical values of the Cragg-Donald test statistic as a function of the maximal bias or maximal size distortion allowable, the researcher can even convert the statistic computed into the limiting bias or size distortion. The null hypothesis is in a metric that is natural and important for the researcher, namely the size distortion or bias that results from weak identification.

The technical reason why macronumerosity isn't a problem here is that weak instrument asymptotics are modelled using the "local to zero" approach, i.e., holding the information in the instruments constant as the sample size goes to infinity.

3 Toward informative misspecification testing

In many cases, small modifications of misspecification tests can produce more informative results than the dichotomous NHMT approach can. We already described how the artificial regression approach of Arellano (1993) provides a natural interpretation of the test results, in contrast to Hausman's (1978) classic misspecification test. In this section, we present a novel example building on White's test for heteroskedasticity, but note that extending to cluster-robust standard error estimation and inference is straightforward using an analogous approach.

3.1 Testing for heteroskedasticity revisited: A proposed approach

We propose the following approach to implementing and interpreting White's test. The basic idea is to put the vector-of-contrasts into an interpretable metric: the difference between a classical and a heteroskedastic-robust estimate of the variance of a single standardized (unit-variance) regressor of interest.

For each regressor of interest x_k:

1. Partial out all other regressors. NB: this step leaves unchanged the corresponding elements of $\widehat{S}_{HC}$ and $\widehat{S}_{OIM}$.
2. Standardize the regressor of interest x_k so that it has zero mean and unit variance. Call this transformed regressor $\widetilde{x}_k$. Call the corresponding standardized coefficient $\widetilde{\beta}_k$

Note that this standardizes x_k after partialling out the other regressors. It therefore differs from the usual standardization, which is "pre-partialling-out".

Since $\widetilde{x}_{ik}$ is mean zero and unit variance,

$$S_{xx} = \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_{ik}^2 = 1 \tag{19}$$

and so the estimator of the asymptotic variance of $\widehat{\widetilde{\beta}}_{OLS}$ is

$$\widehat{Avar(\widehat{\widetilde{\beta}}_{OLS})} = \widehat{S} \tag{20}$$

In other words, by using a standardized regressor we can interpret the estimator of the variance of the score as the estimator of the variance of the OLS coefficient on the standardized regressor.

In the expressions above, the vector Ψ_i becomes the scalar $\tilde{x}_{ik}^2$.[5]

c_n is now a scalar and interpretable as the difference in the estimated variance of the standardised OLS coefficient on the regressor of interest:

$$c_n \equiv \widehat{Avar(\widehat{\widetilde{\beta}})}_{HC} - \widehat{Avar(\widehat{\widetilde{\beta}})}_{OIM} = \frac{1}{n}\sum_{i=1}^{n} (\hat{\varepsilon}_i^2 - \hat{\sigma}^2)\tilde{x}_{ik}^2 \tag{21}$$

The estimators of B simplify a bit:

$$\widehat{B} = \frac{1}{n}\sum_{i=1}^{n} [\hat{\varepsilon}_i^2 - \hat{\sigma}^2]^2 (\tilde{x}_{ik}^2 - 1)^2 \tag{22}$$

$$\widehat{B}_{nR2} = \frac{1}{n}\sum_{i=1}^{n} [\hat{\varepsilon}_{in}^2 - \hat{\sigma}_n^2]^2 \;\; \frac{1}{n}\sum_{i=1}^{n} (\tilde{x}_{ik}^2 - 1)^2 \tag{23}$$

So we can construct, for each regressor, a 95% confidence interval for c_n:

$$\left[\; c_n - 1.96 * \sqrt{\widehat{B}}, \quad c_n + 1.96 * \sqrt{\widehat{B}} \; \right] \tag{24}$$

c_n is the difference in the asymptotic variance of OLS, but it's more traditional to work with the estimated variance, i.e., after dividing by n.

Define

$$d_n \equiv \frac{1}{n} c_n \tag{25}$$

[5] As noted above, a commonly-used method of reducing the dimensionality with the While/Koenker/Breusch-Pagan/Godfrey test is to include just levels of the regressors. The above suggests that if the motivation is size distortion using the classical OLS variance estimator, including just the squares makes more sense.

and we can construct, for each regressor,

$$\left[\; d_n - 1.96 * \frac{1}{n}\sqrt{\widehat{B}}, \quad d_n + 1.96 * \frac{1}{n}\sqrt{\widehat{B}} \; \right] \tag{26}$$

The interval is interpretable as the difference in the estimated variance of the OLS coefficient $\widehat{\widehat{\beta}}_k$ for a standardized regressor.

We could then combine this interval estimate of the variance of the coefficient to compose a very different interval for the OLS coefficient than the traditional confidence interval that uses only the point estimate of the variance. Comparing this new interval to the traditional interval would tell us directly about the sensitivity of inference to an assumption of homoskedasticity. A comparable argument applies to an assumption of zero ICC, and a test for clustering of errors, where one could instead measure the sensitivity of inference to different assumptions about clustering.

4 Conclusions

We extend the critique of null hypothesis significance testing (NHST) to misspecification tests and highlight that there are cases where the null hypotheses misspecification testing (NHMT) paradigm is correct, e.g. in the case of a test for underidentification where failure to reject the null should lead one to discard the estimate altogether. In other settings, the NHST critique applies to misspecification tests, and there is interesting variation in how researchers should think about abandoning a dichotomous decision rule for misspecification testing. We provide one example, but we believe this is a fruitful area for future research.

References

AMRHEIN, V., S. GREENLAND, AND B. MCSHANE (2019): "Scientists rise up against statistical significance," *Nature*, 305–307.

ANDERSON, T. (1951): "Estimating linear restrictions on regression coefficients for multivariate normal distributions," *Annals of Mathematical Statistics*, 327–351.

ANDREWS, I., J. H. STOCK, AND L. SUN (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

ANGRIST, J. AND J.-S. PISCHKE (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 1 ed.

ARELLANO, M. (1993): "On the testing of correlated effects with panel data," *Journal of Econometrics*, 59, 87–97.

CATTANEO, M. D. AND M. JANSSON (2018): "Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency," *Econometrica*, 86, 955–995.

CRAGG, J. G. AND S. DONALD (1993): "Testing Identifiability and Specification in Instrumental Variable Models," *Econometric Theory*, 9, 222–240.

GODFREY, L. (1988): *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*, Econometric Society Monographs No. 16., Cambridge, UK: Cambridge University Press.

HAUSMAN, J. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1271.

HAYASHI, F. (2000): *Econometrics*, Princeton, NJ [u.a.]: Princeton Univ. Press.

KLEIBERGEN, F. AND R. PAAP (2006): "Generalized reduced rank tests using the singular value decomposition," *Journal of Econometrics*, 133, 97–126.

MUNDLAK, Y. (1978): "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46, 69–85.

PAGAN, A. (1990): "Evaluating Models: A Review of L.G. Godfrey Misspecification Tests in Econometrics," *Econometric Theory*, 273–281.

RAMSEY, J. (1969): "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis," *Journal of the Royal Statistical Society: Series B (Methodological)*, 31, 350–371.

SARGAN, D. (1975): *Discussion on Mis-specification*, London: Heinemann, 321–322.

STAIGER, D. AND J. H. STOCK (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

STOCK, J. AND M. YOGO (2005): "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models*, ed. by D. W. Andrews, New York: Cambridge University Press, 80–108.

WASSERSTEIN, R. L. AND N. A. LAZAR (2016): "The ASA Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, 70, 129–133.

WHITE, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–38.

WINDMEIJER, F. (2018): "Testing Over- and Underidentification in Linear Models, with Applications to Dynamic Panel Data and Asset-Pricing Models," Bristol Economics Discussion Papers 18/696, School of Economics, University of Bristol, UK.

WOOLDRIDGE, J. M. (1990): "A Unified Approach to Robust, Regression-Based Specification Tests," *Econometric Theory*, 6, 17–43.