Null Hypothesis Misspecification Testing (NHMT) Revisited: How (Not) to Test Orthogonality Conditions

Mark E Schaffer (Heriot-Watt University, Edinburgh)

Swiss Stata Conference 2025
University of Bern, Switzerland, 21 November 2025

P-values and 'Statistical Significance'

P-values, 'statistical significance', 'null hypothesis significance testing' (NHST)

- Much attention in the applied statistics literature in recent years, most of it critical.
- (Curiously, much less attention to this in the econometrics literature and community.)
- American Statistical Association 2016 "Statement on Statistical Significance and P-Values": "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold." (Wasserstein and Lazar, 2016)
- 2019 *Nature* paper by Amrhein et al. (2019), cosigned by over 800 researchers (including this author): researchers should "retire statistical significance" in favor of more nuanced interpretation.

A typical NHST example

A researcher estimates

$$y_i = x_i\beta + \varepsilon_i$$

usually with some 'controls', and then tests the null hypothesis

$$H_0 : \beta = 0$$

based on the estimated $\hat{\beta}$ and its standard error. If the p-value is less than 5%, the researcher declares victory: β is 'statistically significant' and it's time to write it up and send it off to a journal.

There is a long list of reasons why this is Bad Practice.

A typical NHST example (continued)

NHST: a researcher tests

$$H_0 : \beta = 0$$

based on the estimated $\hat{\beta}$ and its standard error, and if the p-value is less than 5%, the researcher declares victory: β is 'statistically significant'.

First on our list: it very likely doesn't help answer any question of economic interest.

As economists, we almost always want to know the answers to 'How big is the effect?' and 'How precisely is it estimated?' Testing $H_0 : \beta = 0$ answers neither of these questions.

Say the researcher rejects the null:

- What if $\hat{\beta}$ is extremely small but extremely precisely estimated?
- What if $\hat{\beta}$ is very large but the standard error is also huge?

(It is amazing that so many papers with this error still get published.)

A typical NHST example (continued)

What should be done instead?

This is where there is less consensus. An optimist's perspective: there is a lot of good practice to choose from, and the lack of consensus on 'best practice' is not a problem.

A minimalist alternative:

$$y_i = x_i\beta + \varepsilon_i$$

and report a frequentist confidence interval for β. CIs have their own interpretation issues, but they are minor compared to the NHST problem.

NB: the pedagogical challenge is how to teach the concept of "coverage".

I now turn to testing for misspecification.

Testing for misspecification

Austin Nichols and myself, in Nichols and Schaffer (2022), examined misspecification testing from the same perspective as the critical literature on NHST.

We argued that the same problems that face NHST also apply to "null hypothesis misspecification testing" (NHMT).

That paper: Ramsey's RESET test, White's test for heteroskedasticity, tests of instrument relevance, Hausman's misspecification test for fixed vs random effects.

This paper: tests of orthogonality conditions in models estimated by instrumental variables (IV) and the Generalized Method of Moments (GMM).

Testing for misspecification

Terminology alert: "specification testing" is most commonly used now, but the early literature also used 'misspecification testing'. Austin and I prefer the latter, in part because it gives us a nice acronym.

A test of misspecification is a test of the assumptions made by the research with respect to a model's estimation and testing.

1. Misspecification tests are a mainstay of applied econometrics, in both practice and pedagogy.
2. Example: a researcher using a linear IV model cannot get published without reporting tests of identification.
3. Any mainstream econometrics textbook - undergraduate or graduate - that omitted specification tests would be criticized as inadequate and incomplete.

Null hypothesis misspecification testing (NHMT)

Misspecification tests are reported in terms of a test of a null hypothesis that the specified model is correct.

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.

In short: standard pedagogy and practice teaches and employs "null hypothesis misspecification testing" (NHMT).

Curiously, the large literature on NHST, as far as Austin and I could tell, has so far entirely ignored this practice of NHMT.

Surprising: many of the criticisms of dichotomous "reject or not" testing clearly apply equally or even more strongly to misspecification tests.

Example: "macronumerosity" (in homage to Goldberger). With enough data, the researcher rejects the null because they have enough information in the dataset to detect model assumptions that are just approximations.

But how **should** the researcher respond?

Practical steps to improve specification testing

How **should** testing for misspecification be done?

... Wait for it...

It depends on the test.

- Main recommendation: misspecification tests should, if possible, be cast in a metric that is immediately useful and informative to the researcher.
- First choice of metric is that of the parameter of interest. E.g., if a model assumption fails, can this failure be expressed in terms of the size of the bias in the parameter of interest?
- Not always possible in the standard formulation of a test, but may be possible to recast the test in a way that allows this. Example: White's test for heteroskedasticity (see Nichols and Schaffer (2022)).
- And not always necessary. Some tests are, in fact, most naturally interpreted in the NHMT framework. The dictum of "Retire statistical significance" applies only selectively to NHMT.

Testing orthogonality conditions

How should testing for **violations of orthogonality conditions** be done?

Hausman tests are the natural starting point.

In fact, the standard Hausman test for fixed vs random effects is a test of orthogonality conditions, and is easily made robust for the non-i.i.d. case.

But this approach does not extend straightforwardly to tests of orthogonality conditions in linear models:

- Hausman tests coincide with GMM Distance tests only in the special case where they have the same degrees of freedom and the i.i.d. assumption is maintained.
- The Hausman test lacks power to detect violations of certain combinations of orthogonality conditions.
- GMM Distance tests don't have this problem, but artificial regression versions are not easily interpreted because they are not in the metric of the parameters being estimated.

Testing orthogonality conditions

How should testing for violations of orthogonality conditions be done?

Solution: estimate multiple specifications with different combinations of instruments jointly in a pooled estimation, and examining the differences between the estimated $\widehat{\beta}$s.

If a cluster-robust covariance estimator is used when estimating the system, it is easy to construct standard errors and confidence intervals for the differences between these $\widehat{\beta}$s.

System estimation for Hausman tests for inefficient estimators proposed by Weesie (1999) and Creel (2004) (and then mostly forgotten). But they were working in the NHMT paradigm.

But the key advantage to this approach to testing orthogonality conditions is that **the tests are automatically reported in the metric of the parameters of interest**. Easy to assess the consequences of failures of orthogonality conditions and how precisely they're estimated!

Outline

Outline for rest of paper:

- H, J and D statistics in linear IV/GMM models.
- Why the artificial regression versions of J and D are not straightforward to interpret.
- The pooled estimation approach to testing orthogonality conditions.
- IV cross-section example - the Acemoglu et al. (2001) paper on the colonial origins of economic development.

GMM estimator

Model to estimate:

$$y_i = \mathbf{x}_i'\beta + \varepsilon_i$$

where the K regressors $\mathbf{x}_i$ are potentially endogenous.

L potential instruments that satisfy the orthogonality conditions

$$E(\mathbf{z}_i\varepsilon_i) = \mathbf{0}$$

where the $L \times K$ matrix $E(\mathbf{z}_i\mathbf{x}_i')$ is full column rank.

The GMM estimator with weighting matrix $\mathbf{W}_n$ solves

$$\widehat{\beta}_{GMM} = \underset{\widehat{\beta}_0}{\arg\min}\ J_n(\widehat{\beta}_0) \qquad J_n(\widehat{\beta}_0) = n\ \overline{\mathbf{g}}_n(\widehat{\beta}_0)'\mathbf{W}_n\overline{\mathbf{g}}_n(\widehat{\beta}_0)$$

where

$$\overline{\mathbf{g}}_n(\widehat{\beta}_0) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i \qquad \mathbf{g}_i \equiv \mathbf{z}_i\widehat{\varepsilon}_i \equiv \mathbf{z}_i(y_i - \mathbf{x}_i'\widehat{\beta}_{GMM})$$

Feasible Efficient GMM (FEGMM) estimator

The Feasible Efficient GMM estimator is the GMM estimator where the weighting matrix $\mathbf{W}_n$ is $\widehat{\mathbf{S}}^{-1}$. The matrix $\widehat{\mathbf{S}}$ is a consistent estimate of $\mathbf{S} \equiv AVar(\overline{\mathbf{g}}_n)$, the asymptotic variance of $\overline{\mathbf{g}}_n$.

The Feasible Efficient GMM estimator solves

$$\widehat{\beta}_{FEGMM} = \underset{\widehat{\beta}_0}{\arg\min}\ J_n(\widehat{\beta}_0) \qquad J_n(\widehat{\beta}_0) = n\ \overline{\mathbf{g}}_n(\widehat{\beta}_0)'\widehat{\mathbf{S}}^{-1}\overline{\mathbf{g}}_n(\widehat{\beta}_0)$$

Classic Hansen-Sargan J test: value of the minimized efficient GMM objective function $J_n(\widehat{\beta}_{FEGMM})$. If the full model is correctly specified, the value of the minimized GMM objective function using $\widehat{\mathbf{S}}^{-1}$ has a χ^2 distribution with $L - K$ degrees of freedom:

$$J_n(\widehat{\beta}_{FEGMM}) \underset{d}{\to} \chi^2_{L-K}$$

Usual interpretation: a test of the orthogonality conditions $E(\mathbf{z}_i\varepsilon_i) = \mathbf{0}$. Large test statistic suggests failure of one or more orthogonality conditions.

Testing orthogonality conditions

The J test is usually interpreted as a joint test of all the orthogonality conditions $E(\mathbf{z}_i\varepsilon_i) \equiv E(\mathbf{g}_i) = \mathbf{0}$.

Subsets of orthogonality conditions can be tested using the GMM Distance Statistic D.

Setup:

$E(\mathbf{g}_{i,AB}) = \mathbf{0}$	Full set of L orthogonality conditions
$E(\mathbf{g}_{i,A}) = \mathbf{0}$	L_A moment conditions believed to be true
$E(\mathbf{g}_{i,B}) = \mathbf{0}$	$L_B \equiv L - L_A$ moment conditions to be tested

We define two different efficient GMM estimators.

$\widehat{\beta}_{AB}$ uses the full set of orthogonality conditions A and B.

$\widetilde{\beta}_{A}$ uses only orthogonality conditions A.

GMM Distance Statistic D

The GMM Distance statistic:

$$D \equiv J_n(\widehat{\beta}_{AB}) - J_n(\widetilde{\beta}_A)$$

D is the difference between minimized J statistics of the restricted and unrestricted efficient GMM estimators.

Under the null hypothesis $H_0 : E(\mathbf{g}_{i,B}) = \mathbf{0}$, the GMM Distance statistic D has a χ^2 distribution with L_B degrees of freedom:

$$D \underset{d}{\to} \chi^2_{L_B}$$

Standard interpretation of a large D: reject the null hypothesis and conclude that $E(\mathbf{g}_{i,B}) \neq 0$.

Better interpretation: a large D indicates that the two sets of orthogonality conditions are identifying different βs, and a small D suggests that the two sets of orthogonality conditions are identifying the same or similar βs.

Hausman's specification test

This intuition sounds like the intuition behind Hausman's specification test. We'll see in a moment there's a good reason for this.

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. 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.

Important: to construct the vector of contrasts we use the IV estimator, i.e., the Hausman test assumes homoskedasticity.

Hausman's specification test (continued)

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

$$H = n\left(\widehat{\beta}_A - \widehat{\beta}_{AB}\right)' \left(\mathbf{V}(\widehat{\beta}_A) - \mathbf{V}(\widehat{\beta}_{AB})\right)^{-1} \left(\widehat{\beta}_A - \widehat{\beta}_{AB}\right)$$

Under $H_0 : E(\mathbf{g}_{i,B}) = \mathbf{0}$ and the maintained assumption that $E(\mathbf{g}_{i,A}) = \mathbf{0}$, the Hausman statistic H is distributed as $\chi^2(L_B)$, the number of suspect moment conditions being tested.

The problem here is macronumerosity: with real world data, the individual orthogonality conditions are unlikely to be exactly true.

With a large enough sample size and real data, the researcher will reject H_0, even if the vector of contrasts between $\widehat{\beta}_A$ and $\widehat{\beta}_{AB}$ is extremely small in practical terms.

Hausman's specification test (continued)

Much better: interpret the Hausman test in the metric of the test itself, i.e., the parameter estimates.

Easy to do: the Hausman procedure automatically generates a covariance matrix for the vector of contrasts.

In fact, it is automatically reported in the output of Stata's `hausman` command.

Key point: the test with SEs and confidence intervals for the difference in $\hat{\beta}$s.. The researcher can immediately assess how important in practical terms are any differences between $\widehat{\beta}_A$ and $\widehat{\beta}_{AB}$.

GMM Distance Statistic D and Hausman H

The intuition for the interpretation of the GMM Distance Statistic D sounded similar to the interpretation of a Hausman test, but the relationship is closer than that.

Say independence and conditional homoskedasticity hold, so IV and OLS estimators are efficient GMM estimators.

Then it can be shown (see Hayashi (2000), pp. 233-4 and the references therein) that under conditional homoskedasticity and independence, **if** H and D have the same degrees of freedom, **they are numerically equal**.

What does it mean when the degrees of freedom are different?

The Hausman test is designed to find violations of orthogonality conditions via a single contrast of two βs.

When D has more degrees of freedom, it is in effect contrasting all the βs identified by the different instruments. J and D therefore have power to detect violations that H cannot.

J, D, H and NHMT

The drawback with the traditional approach is again NHMT:

- Unrealistic to expect orthogonality conditions to be perfectly satisfied.
- With enough data, even a minor violation that creates a small and economically meaningless but nonzero bias will lead to the rejection of the null that the instruments are valid.
- Noisy estimates may lead to a failure to reject, not because all the instruments are valid but because there is not enough information in the data to detect meaningful violations.

How to proceed?

Problem: drawbacks to traditional approaches to both H and J/D.

Limitations to traditional H

The appeal of the Hausman formulation is that problems with orthogonality conditions are in a readily interpretable metric, namely the coefficient estimates themselves.

But the original Hausman approach is not available for the non-i.i.d. case, and it lacks power to detect certain violations of the orthogonality conditions.

An artificial regression approach to the Hausman test is possible, and can be made robust to non-i.i.d. settings (see e.g. Wooldridge (2002), chapter 6).

But the coefficients on the artificial regressors do not have a convenient interpretation, and the problem of lack of power in certain directions remains.

Limitations to traditional J and D

An artificial regression approach to J/D/GMM-type tests also is available, but it too has drawbacks that make it difficult to use to assess orthogonality failures.

Use the same notation as above, i.e., the $\mathbf{z}_{i,B}$ instruments are to be tested. Also denote the projection matrix using the A instruments by P_{zA}.

Artificial regression version: estimate including $(I - P_{zA})z_{iB}$, i.e., the residuals from projecting the suspect instruments onto the A instruments. A joint test of the significance of the coefficients is equivalent to a GMM-type test of the suspect orthogonality conditions.

The problem here is that the additional regressors do not have an easy interpretation; they are in the metric of the instruments, rather than the regressors. No straightforward interpretation in terms of the βs is available.

Another approach is needed.

Testing orthogonality conditions via pooled regressions

A simple and direct approach: why not estimate the equation of interest separately using different combinations of instruments as desired, and examine directly the difference between the estimated βs via a pooled regression?

This approach was suggested by Weesie (1999) and Creel (2004), who were motivated primarily to generalize the Hausman approach in order to accommodate inefficient as well as efficient estimators, but still working within the NHMT paradigm.

The key payoff to this approach, however, is that it lets us go beyond the "reject" vs "fail to reject" conclusions of NHMT to examine directly how different orthogonality conditions and their failures affect the parameter estimates.

An example

Consider the basic case of $K = 1$, i.e., a single endogenous regressor and no exogenous regressors.

Say also we have three possible instruments: $z_{i,A}$, $z_{i,B}$ and $z_{i,C}$.

The instruments can be used in various combinations; say we use them individually in three separate just-identified estimations, plus all three together in an overidentified estimation.

We can estimate this set of four equations as a system. The simplest way to do this is to stack the data matrices and then estimate the pooled regression using instrumental variables.

The covariance estimator for this pooled regression is just the standard cluster-robust covariance estimator, where we cluster by observation i. No other adjustment needed!

An example (continued)

$$\mathbf{Y} \equiv \begin{bmatrix} \mathbf{y} \\ \mathbf{y} \\ \mathbf{y} \\ \mathbf{y} \end{bmatrix} \qquad \mathbf{X} \equiv I(4) \otimes \mathbf{x} \equiv \begin{bmatrix} \mathbf{x} & 0 & 0 & 0 \\ 0 & \mathbf{x} & 0 & 0 \\ 0 & 0 & \mathbf{x} & 0 \\ 0 & 0 & 0 & \mathbf{x} \end{bmatrix}$$

The instrument set is a $4n \times 6$ data matrix; each block of n observations corresponds to a different IV estimator.

The IV estimator is simply the vector of the four different possible IV estimators.

$$\mathbf{Z} \equiv \begin{bmatrix} \mathbf{z}_A & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{z}_B & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{z}_C & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{z}_A & \mathbf{z}_B & \mathbf{z}_C \end{bmatrix} \qquad \widehat{\mathbf{B}} \equiv \begin{bmatrix} \widehat{\beta}_A \\ \widehat{\beta}_B \\ \widehat{\beta}_C \\ \widehat{\beta}_{ABC} \end{bmatrix}$$

An example (continued)

The IV estimator is simply the vector of the four different possible IV estimators.

$$\mathbf{Z} \equiv \begin{bmatrix} \mathbf{z}_A & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{z}_B & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{z}_C & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{z}_A & \mathbf{z}_B & \mathbf{z}_C \end{bmatrix} \qquad \widehat{\mathbf{B}} \equiv \begin{bmatrix} \widehat{\beta}_A \\ \widehat{\beta}_B \\ \widehat{\beta}_C \\ \widehat{\beta}_{ABC} \end{bmatrix}$$

The error terms in the four equations are of course correlated, but this is easily addressed by using a cluster-robust covariance estimator - clustering on the observation number i - that is robust to within-group correlation as well as heteroskedasticity.

The researcher can then examine potential failures of orthogonality conditions simply by the difference between two estimated βs. Standard errors and confidence intervals for these comparisons are easily obtained.

IV/GMM example: AJR (2001)

Acemoglu-Johnson-Robinson (AJR), "The Colonial Origins of Comparative Development", AER 2001, http://economics.mit.edu/files/4123.

Research question: What is the impact of good economic institutions on the level of economic development in countries?

Model:	where	y_i	GDP per capita in 1995
$y_i = \beta x_i + \varepsilon_i$		x_i	Measure of quality of institutions

Key endogeneity problem: **Reverse causality.** Rich countries can afford to pay for good economic institutions (legal systems, regulation, etc.); poor countries cannot. $E(x_i \varepsilon_i) > 0$ and $\widehat{\beta}_{OLS}$ is biased upwards because

development $\Rightarrow$ good institutions

as opposed to (and in addition to?)

good institutions $\Rightarrow$ development

as in model.

IV/GMM example: AJR (2001)

How does IV/GMM help us solve the reverse causality problem?

Model:	where	y_i	GDP per capita in 1995
$y_i = \beta x_i + \varepsilon_i$		x_i	Measure of quality of institutions

AJR solution: IV, where the excluded instrument z_i is:

$logem_i$ Mortality rates of European colonial settlers (log death rate)

AJR argument: European colonial powers either (1) set up 'little Europes', complete with European settlers, laws, business codes, etc.; or (2) exploit the colony by extracting raw materials etc. and don't build good institutions. If the colony is a dangerous place for Europeans because of diseases (malaria, yellow fever, etc.), the colonial power would follow strategy (2), otherwise it follows (1).

Low settler death rates = colony attractive, build institutions. High death rates = colony not attractive, exploit. After colonial power leaves, any institutions created are still there, and so today $E(logem_i x_i) < 0$.

IV/GMM example: AJR (2001)

Model:	where		
$y_i = \beta x_i + \varepsilon_i$		y_i	GDP per capita in 1995
		x_i	Measure of quality of institutions

Main AJR instrument:

$logem_i$ Mortality rates of European colonial settlers (log death rate)

This instrument got all the attention, but in fact AJR proposed using a number of other instruments. Here we focus on two others:

$euro1900_i$ percentage of the population in 1900 of European descent

$democ1_i$ measure of democracy in the first year of independence

And the question is, do these instruments all identify the same causal effect β of institutions on economic development?

Estimates of the AJR model with different instruments:

Coefficient or diff. in coefs	**Point estimate**	**Standard error**	**95% CI**
OLS (for reference)	0.456	(0.061)	[0.034, 0.574]
$\widehat{\beta}_{ABC}$ (all 3 instruments) $J = 3.20 \quad (p = 0.20)$	0.777	(0.087)	[0.606, 0.948]
$\widehat{\beta}_A$ (IV is *logem4*)	0.905	(0.219)	[0.475, 01.335]
$\widehat{\beta}_B$ (IV is *euro1900*)	0.920	(0.159)	[0.609, 1.231]
$\widehat{\beta}_C$ (*IV is democ1*)	0.383	(0.188)	[0.015, 0.752]
$\widehat{\beta}_{ABC}$ - $\widehat{\beta}_A \quad (p = 0.47)$	-0.128	(0.176)	[-0.473, 0.216]
$\widehat{\beta}_{ABC}$ - $\widehat{\beta}_B \quad (p = 0.24)$	-0.143	(0.121)	[-0.380, 0.094]
$\widehat{\beta}_{ABC}$ - $\widehat{\beta}_C \quad (p = 0.06)$	0.393	(0.209)	[-0.016, 0.803]
$\widehat{\beta}_A$ - $\widehat{\beta}_B \quad (p = 0.95)$	0.015	(0.219)	[-0.444, 0.414]
$\widehat{\beta}_A$ - $\widehat{\beta}_C \quad (p = 0.13)$	0.522	(0.344)	[-0.153, 1.197]
$\widehat{\beta}_B$ - $\widehat{\beta}_C \quad (p = 0.08)$	0.536	(0.295)	[-0.042, 1.115]
$H_0 : \widehat{\beta}_A = \widehat{\beta}_B$ and $\widehat{\beta}_A = \widehat{\beta}_C$ $\chi^2(2) = 3.33 \quad (p = 0.19)$			

Notes:
p-values in parentheses for the differences in estimated $\widehat{\beta}$s are tests of the null hypothesis that the difference is zero. Test statistics are robust to heteroskedasticity. Estimations include a constant and a control for geography.

IV/GMM example: Summary

- The standard NHMT approach would be to look at the J statistic using all 3 instruments, conclude all is probably OK ($J = 3.2$, $p = 0.20$), and move on.
- But if we express the tests of orthogonality conditions in the metric of β, it's very clear what is going on.
- *logem*4 and *euro*1990 are identifying virtually the same β: $\widehat{\beta} \approx 0.9$.
- *democ*1 is identifying a much lower β, lower even than the OLS estimate: $\widehat{\beta} \approx 0.4$.
- But these are not very precise estimates: the standard error for the differences between the estimated $\widehat{\beta}$s is 0.2-0.3.
- In short, we have some weak evidence that one of the instruments is identifying an effect that is different from the others, but nothing that sheds doubt on the overall results.

How to do this in Stata?

Simple: use the `gmm` command.

- `gmm` does multiple-equation GMMM.
- Each equation gets its own set of instruments (exogenous variables, orthogonality conditions).
- Just repeat the equation of interest as many times as desired, each time with a different instrument or set of instruments.
- If you want to engage in NHMT, then `test` for the equality of the coefficient(s) on the endogenous regressor(s).
- But much more informative - use `lincom` and look at the confidence intervals for the differences between the estimated coefficients.

But before we do that, let's review how Stata reports Hausman tests and J tests of overidentification.

IV estimation, just identified, homoskedasticity assumed

This is estimation A: a single instrument (`logem`).

```
. ivregress 2sls logpgp95 lat_abst (avexpr=logem4)

Instrumental-variables 2SLS regression          Number of obs   =         59
                                                Wald chi2(2)    =      33.10
                                                Prob > chi2     =     0.0000
                                                R-squared       =     0.1839
                                                Root MSE        =     .89468

------------------------------------------------------------------------------
    logpgp95 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      avexpr |    .905059   .1995468     4.54   0.000     .5139545    1.296164
    lat_abst |  -.5598326   1.191943    -0.47   0.639    -2.895999    1.776333
       _cons |   2.236234   1.169736     1.91   0.056    -.0564057    4.528874
------------------------------------------------------------------------------
Endogenous: avexpr
Exogenous:  lat_abst logem4

est store H_A
```

IV estimation, overidentified, homoskedasticity assumed

This is estimation B: two instruments (`logem` and `democ1`).

```
. ivregress 2sls logpgp95 lat_abst (avexpr=logem4 democ1)

Instrumental-variables 2SLS regression          Number of obs   =         59
                                                Wald chi2(2)    =      44.31
                                                Prob > chi2     =     0.0000
                                                R-squared       =     0.4661
                                                Root MSE        =     .72365

------------------------------------------------------------------------------
    logpgp95 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      avexpr |   .6924884   .1380628     5.02   0.000     .4218902    .9630865
    lat_abst |   .3004907   .9027615     0.33   0.739    -1.468889    2.069871
       _cons |   3.465063   .8133885     4.26   0.000     1.870851    5.059275
------------------------------------------------------------------------------
Endogenous: avexpr
Exogenous:  lat_abst logem4 democ1

. est store H_AB
```

Hausman and J, homoskedasticity assumed

J and H are identical. But the Hausman output is much more informative!

```
. estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =  6.46585  (p = 0.0110)

. hausman H_A H_AB, sigmamore

                 ---- Coefficients ----
             |      (b)           (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      H_A           H_AB        Difference       Std. err.
-------------+------------------------------------------------------------
      avexpr |     .905059     .6924884        .2125707         .083597
    lat_abst |   -.5598326     .3004907       -.8603233        .3383365
------------------------------------------------------------------------------
                      b = Consistent under H0 and Ha; obtained from ivregress.
       B = Inconsistent under Ha, efficient under H0; obtained from ivregress.

Test of H0: Difference in coefficients not systematic

    chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            =   6.47
Prob > chi2 = 0.0110
```

Hausman and J, homoskedasticity assumed

In fact, the *H* test is simply a Wald test of whether the difference between the A and AB coefficients on `avexpr` is zero:

```
. hausman H_A H_AB, sigmamore

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      H_A          H_AB        Difference       Std. err.
-------------+------------------------------------------------------------
      avexpr |     .905059     .6924884        .2125707         .083597
    lat_abst |   -.5598326     .3004907       -.8603233        .3383365
------------------------------------------------------------------------------
                     b = Consistent under H0 and Ha; obtained from ivregress.
      B = Inconsistent under Ha, efficient under H0; obtained from ivregress.

Test of H0: Difference in coefficients not systematic

    chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            =   6.47
Prob > chi2 = 0.0110


. di (.2125707 / .083597)^2
6.465847
```

Not equivalent when degrees of freedom are different!

In this specification, H will always have one degree of freedom because we are always basing the test on the contrast between estimates of the coefficient on the single endogenous regressor `avexpr`.

But D or J will be based on the degree of overidentification. With 3 instruments instead of 2, J will have 2 degrees of freedom, not one. If we reject, then why? Is there a single instrument that's the culprit?

```
. estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(2) =  6.74023  (p = 0.0344)

. hausman H_A H_AB, sigmamore

Test of H0: Difference in coefficients not systematic

    chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            =   1.17
Prob > chi2 = 0.2786
```

How to test using GMM

Estimating equation is always the same; what differs is the choice of instruments.

Below we set up eq0 to use all 3 instruments, and then 3 different equations, each of which uses one of the 3 instruments on its own.

The `onestep` option combined with the `unadjusted` weighting matrix gives us the same estimates as IV with a standard heteroskedastic-robust covariance matrix.

```
gmm ///
 (eq0: logpgp95 - xb0: avexpr lat_abst _cons) ///
 (eq1: logpgp95 - xb1: avexpr lat_abst _cons) ///
 (eq2: logpgp95 - xb2: avexpr lat_abst _cons) ///
 (eq3: logpgp95 - xb3: avexpr lat_abst _cons), ///
 instruments(eq0: lat_abst logem4 euro1900 democ1) ///
 instruments(eq1: lat_abst logem4) ///
 instruments(eq2: lat_abst euro1900) ///
 instruments(eq3: lat_abst democ1) ///
 onestep winitial(unadjusted, indep)
```

```
------------------------------------------------------------------------------
             |               Robust
             | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
xb0          |
      avexpr |   .7767552   .0871493     8.91   0.000     .6059457    .9475648
    lat_abst |  -.0405572   .6293583    -0.06   0.949    -1.274077    1.192962
       _cons |   2.977933   .5769272     5.16   0.000     1.847176    4.108689
-------------+----------------------------------------------------------------
xb1          |
      avexpr |    .905059   .2194348     4.12   0.000     .4749748    1.335143
    lat_abst |  -.5598326   1.068555    -0.52   0.600    -2.654163    1.534498
       _cons |   2.236234   1.330312     1.68   0.093    -.3711299    4.843598
-------------+----------------------------------------------------------------
xb2          |
      avexpr |   .9197618   .1586769     5.80   0.000     .6087608    1.230763
    lat_abst |   -.619338   .9130987    -0.68   0.498    -2.408979    1.170302
       _cons |   2.151241   .9801403     2.19   0.028     .2302009     4.07228
-------------+----------------------------------------------------------------
xb3          |
      avexpr |     .38327   .1879089     2.04   0.041     .0149754    .7515646
    lat_abst |    1.55197   .9209816     1.69   0.092    -.2531212     3.35706
       _cons |   5.252592   1.085239     4.84   0.000     3.125562    7.379622
------------------------------------------------------------------------------
Instruments for equation eq0: lat_abst logem4 euro1900 democ1 _cons
Instruments for equation eq1: lat_abst logem4 _cons
Instruments for equation eq2: lat_abst euro1900 _cons
Instruments for equation eq3: lat_abst democ1 _cons
```

And we're ready to go!

The `gmm` output has everything we need.

The estimated coefficients are all display with their standard errors and confidence intervals,

To do traditional (uninformative) NHMT, use `test` to test the equality of coefficients across equations. For example, we can test whether the coefficients obtained by using the instruments individually are equal:

```
. test [xb0=xb1=xb2]: avexpr

 ( 1)  [xb0]avexpr - [xb1]avexpr = 0
 ( 2)  [xb0]avexpr - [xb2]avexpr = 0

           chi2(  2) =    2.03
         Prob > chi2 =    0.3621
```

This is essentially the same test as a standard J test of overidentifying restrictions (the J statistic that is robust to heteroskedasticity is 3.20 - very similar).

But don't do NHMT - report CIs instead!

Much more informative is to report confidence intervals using `lincom`. For example:

```
. lincom [xb0]avexp - [xb1]avexp

 ( 1)  [xb0]avexpr - [xb1]avexpr = 0

------------------------------------------------------------------------------
             | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         (1) |  -.1283038   .1759162    -0.73   0.466    -.4730933    .2164857
------------------------------------------------------------------------------

. lincom [xb0]avexp - [xb2]avexp

 ( 1)  [xb0]avexpr - [xb2]avexpr = 0

------------------------------------------------------------------------------
             | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         (1) |  -.1430065   .1208765    -1.18   0.237    -.3799201     .093907
------------------------------------------------------------------------------
```

Open questions

Worth going through the trouble of estimating the overidentified version using efficient GMM? (Fiddly to set up.)

Worth a special program that does this automatically and reports output that looks like this?

```
. overidstack
------------------------------------------------------------------------------
             |               Robust
    logpgp95 | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      avexpr |   .7767552   .0871493     8.91   0.000     .6059457    .9475648
    avexpr_1 |    .905059   .2194348     4.12   0.000     .4749748    1.335143
    avexpr_2 |   .9197618   .1586769     5.80   0.000     .6087608    1.230763
    avexpr_3 |     .38327   .1879089     2.04   0.041     .0149754    .7515646
------------------------------------------------------------------------------

Test of overidentifying restrictions:
 ( 1)  avexpr_1 - avexpr_2 = 0
 ( 2)  avexpr_1 - avexpr_3 = 0

           chi2(  2) =    3.33
         Prob > chi2 =    0.1892
```

Conclusions

The traditional NHMT approach to testing orthogonality conditions suffers from the same drawbacks as the traditional NHST approach to estimation of parameters in general.

When examining potential failures of orthogonality conditions, researchers should go beyond the traditional question "Is there evidence that the orthogonality conditions are violated?" and should ask "By how much are they violated?" and "If there is evidence of any violation, does it matter in practical terms?"

The natural framework for answering these questions is the Hausman test, because it allows the research examine these issues in the natural metric of the parameter estimates themselves and how precisely they are estimated.

Conclusions (continued)

But established tests for orthogonality conditions in linear IV/GMM models are not easily interpreted beyond a yes-or-no rejection of the null hypothesis about the orthogonality of the full set (J) or a subset (D) of instruments.

I propose a simple approach that allows interpretations of these tests in the metric of the parameters: estimate the different specifications (different combinations of instruments) as a system and construct tests, differences, standard errors and confidence intervals in the standard way.

The tests constructed using this approach behave in the same way as standard GMM J and D tests, but have this additional feature of easy interpretation that allow us to answer the questions "By how much are the orthogonality conditions violated?" and "If there is evidence of any violation, does it matter in practical terms?"

Thank you!

Null Hypothesis Misspecification Testing (NHMT) Revisited: How (Not) to Test Orthogonality Conditions

Mark E Schaffer (Heriot-Watt University, Edinburgh)

Swiss Stata Conference 2025
University of Bern, Switzerland, 21 November 2025

References I

Ronald L. Wasserstein and Nicole A. Lazar. The asa statement on p-values: Context, process, and purpose. *The American Statistician*, 70 (2):129–133, 2016. doi: 10.1080/00031305.2016.1154108. URL https://doi.org/10.1080/00031305.2016.1154108.

Valentin Amrhein, Sander Greenland, and Blake McShane. Scientists rise up against statistical significance. *Nature*, (567):305–307, 2019.

Austin Nichols and Mark E. Schaffer. Practical steps to improve specification testing. In Nguyen Ngoc Thach, Doan Thanh Ha, Nguyen Duc Trung, and Vladik Kreinovich, editors, *Prediction and Causality in Econometrics and Related Topics*, pages 75–88. Springer International Publishing, 2022. ISBN 978-3-030-77094-5. doi: 10.1007/978-3-030-77094-5_8. URL https://doi.org/10.1007/978-3-030-77094-5_8.

References II

Jeroen Weesie. Seemingly unrelated estimation and the cluster-adjusted sandwich estimator. *Stat Technical Bulletin*, STB-52:34–47, 1999. URL `http://www.stata.com/info/products/stb/journals/stb52.html`.

Michael Creel. Modified hausman tests for inefficient estimators. *Applied Economics*, 36(21):2373–2376, 2004. doi: 10.1080/0003684042000291281. URL `https://doi.org/10.1080/0003684042000291281`.

Daron Acemoglu, Simon Johnson, and James A. Robinson. The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5):1369â1401, December 2001. doi: 10.1257/aer.91.5.1369. URL `https://www.aeaweb.org/articles?id=10.1257/aer.91.5.1369`.

J.A. Hausman. Specification Tests in Econometrics. *Econometrica*, 46(6): 1251–1271, 1978.

References III

Fumio Hayashi. *Econometrics*. Princeton Univ. Press, Princeton, NJ [u.a.], 2000. ISBN 0691010188.

Jeffrey M. Wooldridge. *Econometric Analysis of Cross-Section and Panel Data*. MIT Press, 2002.