econometrics

Article

A New Matrix Statistic for the Hausman Endogeneity Test under Heteroskedasticity

Alecos Papadopoulos

Department of Economics, Athens University of Economics and Business, TK 10434 Athens, Greece; papadopalex@aueb.gr

Abstract: We derive a new matrix statistic for the Hausman test for endogeneity in cross-sectional Instrumental Variables estimation, that incorporates heteroskedasticity in a natural way and does not use a generalized inverse. A Monte Carlo study examines the performance of the statistic for different heteroskedasticity-robust variance estimators and different skedastic situations. We find that the test statistic performs well as regards empirical size in almost all cases; however, as regards empirical power, how one corrects for heteroskedasticity matters. We also compare its performance with that of the Wald statistic from the augmented regression setup that is often used for the endogeneity test, and we find that the choice between them may depend on the desired significance level of the test.

Keywords: Hausman test; heteroskedasticity; endogeneity; instrumental variables; generalized matrix inverse; Wald statistic

JEL Classification: C12; C26; C21

Citation: Papadopoulos, Alecos. 2023. A New Matrix Statistic for the Hausman Endogeneity Test under Heteroskedasticity. *Econometrics* 11: 23. https://doi.org/10.3390/econometrics11040023

Academic Editor: Ryo Okui

Received: 10 August 2023
Revised: 28 September 2023
Accepted: 2 October 2023
Published: 10 October 2023

1. Introduction

The Hausman family of specification tests was introduced by Hausman (1978) and it has seen unabated use in econometrics ever since. Amini et al. (2012) detail its wide reach and different implementations for panel data, while in a cross-sectional setting, the test has often been used to test for regressor endogeneity.

In the cross-sectional setting, the test statistic is formally based on a "vector of contrasts", the difference of two estimators, where under the null hypothesis, both are consistent and the one is also efficient, while under the alternative, only one is consistent. This form of the test uses a variance expression that is often singular, requiring a generalized inverse. To bypass the singularity of the variance matrix an "augmented regression" approach has been developed, linking the test with its precursors (Durbin 1954; Wu 1973, 1974).[1]

The efficiency of one of the estimators under the null hypothesis has a very convenient consequence: the variance of the difference of the two estimators equals the difference of their variances; thus, we do not have to compute covariances. However, when heteroskedasticity is present (and it is expected to exist regularly in cross sectional studies), this helpful simplification is no longer valid. Adkins et al. (2012) have examined in great detail this endogeneity test in situations of heteroskedasticity, and they take the augmented regression route to formulate the various test variants that they implement.

In this study, we push a known result in the literature to its conclusion, and we arrive at a new matrix Hausman statistic for an endogeneity test. This new statistic is a useful additional tool to have, for the following reasons: it handles heteroskedasticity in a natural way; it could be a more familiar tool to use for researchers that are accustomed to using matrix algebra and forms; compared to the original form of the Hausman statistic, it does not use generalized inverses. In fact, if the matrix involved is not invertible, it reflects the existence of perfect collinearity between some instruments and some endogenous regressors, which invalidates the instruments. Finally, in Monte Carlo simulations that we will present, it performed better than the "augmented regression" test in terms of power, when executing the test at the 10% significance level.

2. The Matrix Hausman Statistic for Testing Endogeneity

We follow the notation of Adkins et al. (2012). We consider the linear regression model $\mathrm{y} = X\beta + \mathrm{u}$. The vectors y and u are $n \times 1$, and u is assumed to be zero-mean. The regressor matrix is partitioned $X = [X_1\ X_2]$. X_1 is an $n \times K_1$ submatrix of regressors thought to be endogenous (so not orthogonal to the error term) while X_2 is an $n \times K_2$ submatrix of exogenous regressors (or "internal instruments"). The unknown of interest is the vector β. We have available $\Lambda_1 \geq K_1$ "external instruments" collected in matrix Z_1 and the full instruments matrix is $Z = [Z_1\ X_2]$. We write the orthogonal projection matrix $P_z = \mathrm{Z}(\mathrm{Z}'\mathrm{Z})^{-1}Z'$ and the residual-maker (or annihilator) matrix $M_x = I_n - P_x$, with I_n being the $n \times n$ Identity matrix. The subscript in P and M determines which collection of variables we use in each case. These matrices are symmetric and idempotent. We write $\hat{\mathrm{u}} = M_x\mathrm{y}$ for the residuals from the Ordinary Least Squares regression (OLS). We write $P_zX \equiv \widehat{X} = [\widehat{X}_1\ X_2]$ for the linear projection of X on the columns of Z (the "fitted values"). Note that $P_zX_2 = X_2$ because X_2 belongs in the column space of P_z.

The OLS estimator for β is $\hat{\beta}_{OLS} = (X'X)^{-1}X'\mathrm{y}$ while the benchmark Instrumental Variables (IV) estimator when instruments are more in number than endogenous regressors is $\hat{\beta}_{IV} = (\widehat{X}'\widehat{X})^{-1}\widehat{X}'\mathrm{y}$ (two-stage least-squares). The basic expression of the Hausman statistic under homoskedasticity of the error term (with the OLS estimate of the error variance $\hat{\sigma}_u^2$) is

$$(\hat{\sigma}_u^2)^{-1} \cdot (\hat{\beta}_{IV} - \hat{\beta}_{OLS})'\Big((\widehat{X}'\widehat{X})^{-1} - (X'X)^{-1}\Big)^{-1}(\hat{\beta}_{IV} - \hat{\beta}_{OLS}) \tag{1}$$

This is the statistic where we may encounter trouble in inverting the middle matrix, which, moreover, may not even be positive definite in finite samples. It may render the test inapplicable, or necessitate the use of a generalized inverse instead.

To bypass this issue, while simultaneously accounting for heteroskedasticity, we start by noting that the core of the statistic for the Hausman test is the difference

$$\begin{aligned}
\hat{\beta}_{IV} - \hat{\beta}_{OLS} &= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'\mathrm{y} - (X'X)^{-1}X'\mathrm{y} \\
&= (X'P_zX)^{-1}X'P_z\mathrm{y} - (X'X)^{-1}X'\mathrm{y} \\
&= (X'P_zX)^{-1}\Big[X'P_z\mathrm{y} - X'P_zX(X'X)^{-1}X'\mathrm{y}\Big] \\
&= (X'P_zX)^{-1}X'P_z\Big[I_n - X(X'X)^{-1}X'\Big]\mathrm{y} \\
&= (X'P_zX)^{-1}X'P_zM_x\mathrm{y} \\
&= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'\hat{\mathrm{u}}.
\end{aligned} \tag{2}$$

We have used the fact that P_z, M_x are symmetric and idempotent and that $M_x\mathrm{y} = M_x\mathrm{u} = \hat{\mathrm{u}}$. Result (2) is known in the literature. For example, Greene (2012, p. 276) arrives at it, but he does not go further. Also Adkins et al. (2012) actually start with it (their Equation (1)), but then they use the augmented regression approach to proceed. Also, later in their paper, when they re-purpose the "weak vs. strong instruments" test of Hahn et al. (2011), they "directly estimate the asymptotic covariance matrix of the contrast", but the expression they give is inconveniently long, since this variance can be nicely compacted, as we will show. To our knowledge, the result in Equation (2) has not been pursued to the very end for the construction of a Hausman statistic and test, and this is what we will do here.

The null hypothesis of the Hausman test is that the two estimators converge to the same probability limit (plim):

$$\mathrm{H}_0 : \mathrm{plim}(\hat{\beta}_{IV} - \hat{\beta}_{OLS}) = 0.$$

To examine this hypothesis we consider the limiting distribution of the scaled difference and its variance, which, under the null hypothesis and given Equation (2), is

$$\mathrm{Avar}\left[\sqrt{n}(\hat{\beta}_{IV} - \hat{\beta}_{OLS})\right] \equiv V = \mathrm{plim}\left[(n^{-1}\widehat{X}'\widehat{X})^{-1} S (n^{-1}\widehat{X}'\widehat{X})^{-1}\right]. \tag{3}$$

The middle matrix is $S = \mathrm{plim}(n^{-1}\widehat{X}'\hat{\mathrm{u}}\hat{\mathrm{u}}'\widehat{X})$. We can then formulate a theoretical statistic for the endogeneity test,

$$q = n \cdot (\hat{\beta}_{IV} - \hat{\beta}_{OLS})' V^{-} (\hat{\beta}_{IV} - \hat{\beta}_{OLS}) \xrightarrow[\mathrm{H}_0]{d} \chi^2_{K_1}. \tag{4}$$

Here, V^{-} denotes a generalized inverse of V.[2] Combining Equation (2) and (3) with (4), we arrive at the following statistic feasible to compute, for some consistent estimator $\hat{S}$,

$$\hat{q} = n^{-1} \cdot \hat{\mathrm{u}}'\widehat{X}(\widehat{X}'\widehat{X})^{-1}\left[(\widehat{X}'\widehat{X})^{-1}\hat{S}(\widehat{X}'\widehat{X})^{-1}\right]^{-}(\widehat{X}'\widehat{X})^{-1}\widehat{X}'\hat{\mathrm{u}}. \tag{5}$$

We show in Appendix A that a generalized inverse of the middle matrix is $(\widehat{X}'\widehat{X})\hat{S}^{+}(\widehat{X}'\widehat{X})$, where $\hat{S}^{+}$ denotes the Moore–Penrose generalized inverse. Inserting this in the expression for $\hat{q}$, we can simplify,

$$\hat{q} = n^{-1} \cdot \hat{\mathrm{u}}' \, \widehat{X} \, \hat{S}^{+} \, \widehat{X}'\hat{\mathrm{u}}. \tag{6}$$

Next, because $\widehat{X}$ includes the submatrix X_2, which is by construction orthogonal to the OLS residuals $\hat{\mathrm{u}}$, we obtain that

$$S = \begin{bmatrix} Q_{K_1 \times K_1} & \mathbf{0}_{K_1 \times K_2} \\ \mathbf{0}_{K_2 \times K_1} & \mathbf{0}_{K_2 \times K_2} \end{bmatrix}, \qquad Q = \mathrm{plim}(n^{-1}\widehat{X}_1'\hat{\mathrm{u}}\hat{\mathrm{u}}'\widehat{X}_1). \tag{7}$$

We show in Appendix B that

$$S^{+} = \begin{bmatrix} Q^{-1}_{K_1 \times K_1} & \mathbf{0}_{K_1 \times K_2} \\ \mathbf{0}_{K_2 \times K_1} & \mathbf{0}_{K_2 \times K_2} \end{bmatrix}. \tag{8}$$

We have managed to eliminate the generalized inverse and to use a proper inverse.[3] What remains now is to find a consistent estimator for the Q matrix. Decomposing the OLS residuals, we have

$$\widehat{X}_1'\hat{\mathrm{u}}\hat{\mathrm{u}}'\widehat{X}_1 = \widehat{X}_1' \, M_x \, \mathrm{uu}' \, M_x \, \widehat{X}_1.$$

This matrix expression, where we sandwich the outer product of the error vector, may look familiar to those acquainted with the heteroskedasticity-robust estimation literature and one could expect that we could now use the squared residuals in place of uu' to estimate Q. However, there is an issue: the matrix M_x is $n \times n$, growing in both dimensions as the sample size increases. So it is not clear that the related proof strategy of White (1980) is applicable here. Nevertheless, by drilling down even more, we arrive, in Appendix C, at an expression that contains matrix products with finite dimensions. Thus, we can indeed apply this substitution, which provides a consistent estimator for Q as

$$\hat{Q} = n^{-1}\widehat{X}_1' \, M_x \, \hat{\Omega}_0 \, M_x\widehat{X}_1, \qquad \hat{\Omega}_0 = \mathrm{diag}\{\hat{u}_i^2\}. \tag{9}$$

This indeed looks like a "White" estimator of a heteroskedastic covariance matrix. The expression is valid under the formal assumptions stated in White (1980), which we do not repeat here for brevity.

Equation (9) allows us also to conclude that the matrix Q *must* be invertible; otherwise, at least one component of the instruments matrix is not valid.

This can be shown in the following way: Let $\mathrm{x}_{1j}, j = 1, \ldots, K_1$ be a column of X_1, the submatrix with the endogenous regressors. If $P_z\mathrm{x}_{1j} = \hat{\mathrm{x}}_{1j} = \mathrm{x}_{1j}$, we will have $M_x\hat{\mathrm{x}}_{1j} = \mathbf{0}$ so $\widehat{X}_1' \, M_x$ will have a column of zeros and Q will be singular. However, $P_z\mathrm{x}_{1j} = \mathrm{x}_{1j}$ implies that x_{1j} belongs to the column space of the instruments matrix Z, meaning that it is an *exact* linear combination of the columns of Z. However, if this were the case, it necessarily implies that at least one of the instruments would be correlated with the error term, and so Z too would suffer from endogeneity. Namely, if $\mathrm{x}_{1j} = \sum_{\ell}^{\Lambda_1+K_2} a_\ell \mathrm{z}_\ell$ while $E(\mathrm{x}_{1j}'\mathrm{u}) \neq 0$, we will have $E\left(\mathrm{u}' \sum_{\ell}^{\Lambda_1+K_2} a_\ell \mathrm{z}_\ell\right) \neq 0$.

Therefore, using a proper inverse here also serves the function of alerting us that such exact linear dependence exists between instruments and endogenous regressors, if the matrix Q proves to be non-invertible. In such a case, executing the endogeneity test using a generalized inverse would be wrong; one first has to somehow correct the instruments matrix to restore the validity of the instruments.[4]

Lastly, using these results in the expression for $\hat{q}$, and again the fact that $\hat{\mathrm{u}}' \widehat{X} = \left[\hat{\mathrm{u}}' \widehat{X}_1 \,:\, \mathbf{0}\right]$, we arrive at the final expression for the heteroskedasticity-robust matrix Hausman statistic (where we have also canceled out the n^{-1} factors),

$$\hat{q}_{het} = \hat{\mathrm{u}}' \widehat{X}_1 \left[\widehat{X}_1' M_x \hat{\Omega}_0 M_x \widehat{X}_1\right]^{-1} \widehat{X}_1' \hat{\mathrm{u}} \xrightarrow[\mathrm{H}_0]{d} \chi^2_{K_1}, \qquad \hat{\Omega}_0 = \mathrm{diag}\{\hat{\mathrm{u}}_i^2\}. \tag{10}$$

Computing this statistic first requires running OLS estimation on the original model to obtain the residuals $\hat{\mathrm{u}}$ and their squares for $\hat{\Omega}_0$, and then using the matrices P_z and X_1 (since $\widehat{X}_1 = P_z X_1$), and also the matrix M_x. The matrices P_z and M_x are of dimension $n \times n$; thus, for very large samples, they may be taxing for the software (although they will be used just once in an actual applied study; it is simulation studies that may be considerably slowed down when using them). If one wishes to avoid them, to obtain $\widehat{X}_1$, we can run OLS regressions for the columns of X_1 on Z, and also, to compute $M_x \widehat{X}_1$ we can run regressions of $\widehat{X}_1$ on X and obtain the resulting residual series.

The statistic can be used also under the assumption of homoskedasticity, in which case it becomes

$$\hat{q}_{hom} = (\hat{\sigma}_u^2)^{-1} \hat{\mathrm{u}}' \widehat{X}_1 \left[\widehat{X}_1' M_x \widehat{X}_1\right]^{-1} \widehat{X}_1' \hat{\mathrm{u}} \xrightarrow[\mathrm{H}_0]{d} \chi^2_{K_1}. \tag{11}$$

To increase power, one should use the error variance estimator from the OLS regression.

Equations (10) and (11) are the main theoretical contribution of this study. We have exploited the expression for the vector of contrasts in terms of projected regressors and OLS residuals, and we have arrived at a matrix Hausman statistic that incorporates possible heteroskedasticity in a natural way, it has a compact form, it does not use generalized inverses, and it guards against invalid instruments.

In the next section we present results from a simulation study to examine the performance of this matrix Hausman statistic, looking also into the variants that have been proposed for $\hat{\Omega}$ in an attempt to improve finite-sample performance. A recent overview and Monte Carlo study for these "HCx" estimators for heteroskedastic variance matrices can be found in MacKinnon (2013).

3. Monte Carlo Study

3.1. Description

We constructed a data generation process (DGP) with a constant term, one exogenous variable, two "suspected" endogenous variables and three external valid instruments. We considered a case where the DGP includes an unobservable covariate uncorrelated with the regressors (so here, OLS is consistent even if this variable is not included in the regressor matrix), and one where it is correlated (so there is endogeneity and OLS is inconsistent). The first case serves to examine the empirical size of the test, while the second provides information about the power of the test. The technical details of the Monte Carlo study are presented in Appendix D.

We created four scenarios as regards heteroskedasticity of the error term: homoskedasticity, heteroskedasticity with the error variance randomly changing per observation independently of the regressors, "group-wise" heteroskedasticity, where the error variance takes only three distinct values with equal probability, again independent of the regressors, and finally, a "random-coefficients" model, which leads to the error variance being a function of the regressors, without this affecting mean-independence. We considered sample sizes $n = 50, 75, 100, 200$, and in each case, we executed 10,000 repetitions. In all cases, we initiated the random number generator by the same seed. This has two consequences: first, for a given sample size all scenarios have identical series for the observable variables, and they differ only with respect to the endogeneity/heteroskedasticity aspect. Second, for each scenario, as we increase the sample size the previous generated values are fully part of the larger sample. In this way, we mimic the accumulation of data rather than the availability of independent larger data sets.

As regards the statistic, we used both its homoskedastic variant (i.e., assuming, correctly or not, that the true error is homoskedastic), as well as the four best-known alternatives for estimation of heteroskedastic variance matrices, $\text{HCx}, \text{x} = 0, 1, 2, 3$, as these are defined in MacKinnon (2013): writing h_{ii} for the diagonal element of the projection matrix P_x, we have

$$\begin{aligned}
\text{HC0}\ &: \hat{\Omega}_0 = \text{diag}\{\hat{\text{u}}_i^2\}, \\
\text{HC1}\ &: \hat{\Omega}_1 = \frac{n}{n-k}\text{diag}\{\hat{\text{u}}_i^2\}, \\
\text{HC2}\ &: \hat{\Omega}_2 = \text{diag}\{\hat{\text{u}}_i^2/(1-h_{ii})\}, \\
\text{HC3}\ &: \hat{\Omega}_3 = \text{diag}\{\hat{\text{u}}_i^2/(1-h_{ii})^2\}.
\end{aligned}$$

Note that k is the number of regressors each time. For our matrix statistic, the number of regressors is $k = K_1 + K_2$, where $K_1 = K_2 = 2$.

3.2. Comparative Performance of the Variants of the Matrix Hausman Statistic

Here, we assess how the statistic performs in terms of empirical size and power, as we change the heteroskedasticity correction. We do not compare it with other forms of the Hausman test, because we want first to determine whether it has an acceptable performance (empirical size close to nominal, power rising fast with the sample size). If it performs acceptably, then a case arises to compare it to other forms of the Hausman test. We present the results in Table 1, which relates to testing at the 5% significance level.

We have the following main observations: first, the behavior of the Hausman matrix statistic as regards empirical size is rather stable, across different true skedastic scenarios as well as across different HCx ways to incorporate the possible heteroskedasticity. In fact, it performs acceptably in relation to the size of the test, even if we ignore the possible presence of heteroskedasticity and use (11) instead of (10). Second, for the various HCx variants to account for heteroskedasticity, the empirical size monotonically falls as we increase the strength of the finite-sample correction that we apply. Results for testing at the 10% significance level (available upon request) show a similar behavior in relation to empirical size.

As regards empirical power, the choice of the heteroskedastic variant for $\hat{\Omega}$ matters even more, for small sample sizes. Power also deteriorates monotonically and visibly, while the highest power is achieved when we use the homoskedastic variant (where the test is slightly conservative).[5] Overall, for testing at the 5% significance level, it appears that the prudent thing to do when applying this statistic, is to use its HC0 heteroskedastic formula. When testing at the 10% significance level, power increases visibly. For example, under conditional heteroskedasticity, the power at 10% for sample sizes $n = 50, 75$ tends to be higher by a factor of 1.2 to almost 1.9, i.e., almost double the power at the 5% significance level, all else being equal. For the 10% significance level therefore, it appears best to use the homoskedastic variant of the matrix statistic, Equation (11).

Table 1. Monte Carlo simulation study. Empirical Size and Power of the matrix Hausman statistic. Nominal size: 5%.

	Skedastic Scenario	**Homoskedasticity**		**Random**		**Group-Wise**		**Conditional**	
n	**Robust Estimation**	**Size**	**Power**	**Size**	**Power**	**Size**	**Power**	**Size**	**Power**
50	Homoskedastic	4.69	49.05	4.77	48.43	4.86	47.23	5.50	38.57
	HC0	5.54	41.09	5.55	40.67	5.45	40.11	5.81	32.04
	HC1	4.30	35.06	4.23	34.64	3.96	34.22	4.03	26.53
	HC2	4.03	33.00	4.00	32.22	3.67	32.11	3.59	24.30
	HC3	2.77	24.93	2.70	24.46	2.23	24.10	2.18	17.36
75	Homoskedastic	4.50	71.74	4.61	71.37	4.78	70.51	5.53	57.97
	HC0	5.21	64.98	5.30	64.36	5.35	63.60	5.61	51.21
	HC1	4.37	61.44	4.43	60.94	4.39	60.08	4.53	47.26
	HC2	4.15	59.41	4.09	58.97	4.16	58.06	4.13	45.04
	HC3	3.18	53.43	3.12	52.77	3.15	51.41	2.91	38.67
100	Homoskedastic	4.62	86.16	4.70	85.31	4.80	84.96	5.76	73.35
	HC0	5.02	81.33	5.23	80.72	5.17	80.25	5.05	66.79
	HC1	4.51	79.55	4.34	78.59	4.53	78.39	4.47	64.38
	HC2	4.34	78.23	4.16	77.38	4.24	77.04	4.15	62.46
	HC3	3.43	74.23	3.36	73.25	3.58	72.65	3.32	57.35
200	Homoskedastic	4.90	99.50	4.85	99.45	4.47	99.31	6.08	96.79
	HC0	5.19	99.13	5.00	99.02	4.86	98.95	5.26	94.74
	HC1	4.86	99.06	4.81	98.98	4.50	98.87	4.97	94.34
	HC2	4.70	98.94	4.75	98.85	4.36	98.73	4.79	93.77
	HC3	4.27	98.53	4.29	98.43	3.95	98.32	4.28	92.67

The finding that ignoring heteroskedasticity while it exists may lead to better-performing tests should not be surprising for small samples. To account for heteroskedasticity, we use additional estimated quantities, the OLS residuals *individually* and this should be expected to negatively affect statistical power in the context of a small sample.

3.3. Comparison with the Wald Statistic from the Augmented Regression Approach

Using the exact same simulated data sets, we have also computed the Wald statistic coming from the augmented regression setup to test for endogeneity. Here, we first regress the suspected endogenous regressors X_1 on the full instrument matrix Z, we obtain the residuals $M_z X_1$ and we include these residuals in an augmented regressor matrix $X_A = [X_1 : X_2 : M_z X_1]$.[6] We run an OLS regression of the dependent variable on X_A, and we compute a Wald test for the coefficients of the regressors in the submatrix $M_z X_1$.[7]

In the interest of space, we do not report the full results here. The Wald statistic clearly has a size problem for these small samples: it tends to over-reject the correct null hypothesis, sometimes even having empirical size nearly double the nominal one (for both 5% and 10% nominal significance levels). The over-rejection becomes less than one percentile across skedastic scenarios, only with the HC3 heteroskedasticity correction.[8] In Table 2, we report the performance for this statistic for testing at the 5% significance level, and we repeat the performance metrics for our matrix statistic with the HC0 formula from Table 1.

The Wald statistic appears to have an advantage as regards power, even though some of this advantage will be lost due to the correction for the slightly oversized test. However, the picture changes if we want to test at the 10% significance level. Here, it is our matrix statistic (which moreover assumes homoskedasticity) that has the advantage in terms of power, as is shown in Table 3.

Overall, no statistic dominates the other, and for sample sizes larger than $n = 200$ the two are essentially equivalent in terms of size and power. For smaller samples, the desired significance level of the test can be our guide in order to choose between them.

Table 2. Comparison in empirical size and power of the matrix Hausman statistic vs. the Wald statistic from the augmented regression setup. Nominal size: 5%.

	Skedastic Scenario	**Homoskedasticity**		**Random**		**Group-Wise**		**Conditional**	
n	**Statistic**	**Size**	**Power**	**Size**	**Power**	**Size**	**Power**	**Size**	**Power**
50	$\hat{q}_{het}$-HC0	5.54	41.09	5.55	40.67	5.45	40.11	5.81	32.04
	Wald-HC3	5.76	45.17	5.74	44.84	5.76	44.21	5.64	34.93
75	$\hat{q}_{het}$-HC0	5.21	64.98	5.30	64.36	5.35	63.60	5.61	51.21
	Wald-HC3	5.57	68.42	5.45	67.76	5.53	67.53	5.43	52.96
100	$\hat{q}_{het}$-HC0	5.02	81.33	5.23	80.72	5.17	80.25	5.05	66.79
	Wald-HC3	5.40	83.62	5.29	82.63	5.39	82.56	5.31	67.84
200	$\hat{q}_{het}$-HC0	5.19	99.13	5.00	99.02	4.86	98.95	5.26	94.74
	Wald-HC3	5.36	99.38	5.18	99.31	4.87	99.16	5.39	94.80

Table 3. Comparison in empirical size and power of the matrix Hausman statistic vs. the Wald statistic from the augmented regression setup. Nominal size: 10%.

	Skedastic Scenario	**Homoskedasticity**		**Random**		**Group-Wise**		**Conditional**	
n	**Statistic**	**Size**	**Power**	**Size**	**Power**	**Size**	**Power**	**Size**	**Power**
50	$\hat{q}_{hom}$	9.72	62.89	9.63	62.31	10.14	60.37	10.60	51.81
	Wald-HC3	10.16	57.35	10.15	56.43	10.44	55.14	9.91	45.69
75	$\hat{q}_{hom}$	9.79	82.04	9.79	81.41	9.95	80.93	11.05	70.08
	Wald-HC3	10.11	77.89	10.09	77.51	10.18	77.27	9.95	64.43
100	$\hat{q}_{hom}$	9.77	92.14	9.74	91.68	10.26	91.34	11.23	82.55
	Wald-HC3	10.07	90.45	10.02	89.91	10.26	89.51	10.10	78.26
200	$\hat{q}_{hom}$	9.96	99.80	9.81	99.74	9.96	99.75	11.60	98.45
	Wald-HC3	9.90	99.77	9.93	99.66	10.33	99.62	10.23	97.33

Funding: This research received no external funding.

Data Availability Statement: No data were used for this study.

Conflicts of Interest: The author declare no conflicts of interest.

Appendix A

We argue that a generalized inverse of $(\widehat{X}'\widehat{X})^{-1}\hat{S}(\widehat{X}'\widehat{X})^{-1}$ is $(\widehat{X}'\widehat{X})\hat{S}^{+}(\widehat{X}'\widehat{X})$.

A generalized inverse A^{-} of matrix A satisfies $A\,A^{-}\,A = A$. Setting for compactness $(\widehat{X}'\widehat{X})^{-1} \equiv C^{-1}$ and $A = C^{-1}\hat{S}C^{-1}$, our candidate generalized inverse is $A^{-} = C\hat{S}^{+}C$. We have

$$\begin{aligned} A\,A^{-}\,A &= [C^{-1}\,\hat{S}\,C^{-1}]\,[C\,\hat{S}^{+}\,C][C^{-1}\,\hat{S}\,C^{-1}] \\ &= C^{-1}\,\hat{S}\,\hat{S}^{+}\,\hat{S}\,C^{-1} \\ &= C^{-1}\,\hat{S}\,C^{-1} = A, \end{aligned}$$

which is what we wanted to show. $\hat{S}\,\hat{S}^{+}\,\hat{S} = \hat{S}$ holds because the Moore–Penrose inverse satisfies this condition, among others.

Appendix B

We argue that

$$S = \begin{bmatrix} Q_{K_1\times K_1} & \mathbf{0}_{K_1\times K_2} \\ \mathbf{0}_{K_2\times K_1} & \mathbf{0}_{K_2\times K_2} \end{bmatrix} \implies S^{+} = \begin{bmatrix} Q^{-1}_{K_1\times K_1} & \mathbf{0}_{K_1\times K_2} \\ \mathbf{0}_{K_2\times K_1} & \mathbf{0}_{K_2\times K_2} \end{bmatrix}.$$

In order for a matrix A^+ to be indeed the unique Moore–Penrose pseudo-inverse of matrix A, it must satisfy four conditions:

1. $AA^+A = A$
2. $A^+AA^+ = A^+$
3. $(AA^+)' = AA^+$
4. $(A^+A)' = A^+A$.

Note that Q is symmetric. For condition 1, we have

$$\begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} Q^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} I_{K_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

For condition 2, we have

$$\begin{bmatrix} Q^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} Q^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} I_{K_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} Q^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} Q^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

For condition 3, we have

$$\left(\begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} Q^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right)' = \begin{bmatrix} I_{K_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} Q^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

and for condition 4, we have analogously

$$\left(\begin{bmatrix} Q^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right)' = \begin{bmatrix} I_{K_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} Q^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \cdot \begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Appendix C

We want to find a consistent estimator for

$$Q = \text{plim}(n^{-1}\widehat{X}_1'\hat{\mathrm{u}}\hat{\mathrm{u}}'\widehat{X}_1).$$

We have

$$\widehat{X}_1'\hat{\mathrm{u}} = X_1'\,P_z\,\hat{\mathrm{u}} = X_1'\,Z(Z'Z)^{-1}Z'\,\hat{\mathrm{u}} = X_1'\,Z(Z'Z)^{-1}\begin{bmatrix} Z_1' \\ X_2' \end{bmatrix}\hat{\mathrm{u}}.$$

However, $X_2'\,\hat{\mathrm{u}} = \mathbf{0}$; thus, carrying out the multiplications and including the sample size as a scaling factor, we arrive at

$$Q = \text{plim}\left\{ \left(n^{-1}X_1'Z\right)(n^{-1}Z'Z)^{-1}\begin{bmatrix} n^{-1}Z_1'\hat{\mathrm{u}}\hat{\mathrm{u}}'Z_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}(n^{-1}Z'Z)^{-1}\left(n^{-1}Z'X_1\right) \right\}.$$

The standard regularity conditions are assumed to hold, and so the matrices that sandwich the middle one are well defined and converge to a finite probability limit. Focusing on the middle one, we have

$$\hat{\mathrm{u}}\hat{\mathrm{u}}' = M_x\mathrm{u}\mathrm{u}'M_x = (I_n - P_x)\mathrm{u}\mathrm{u}'(I_n - P_x) = \mathrm{u}\mathrm{u}' - \mathrm{u}\mathrm{u}'P_x - P_x\mathrm{u}\mathrm{u}' + P_x\mathrm{u}\mathrm{u}'P_x.$$

Further, $Z_1'P_x = Z_1'\,X\,(X'X)^{-1}X'$ and $P_xZ_1 = X\,(X'X)^{-1}X'Z_1$; thus, adding scaling factors again, we have

$$\begin{aligned} n^{-1}Z_1'\hat{\mathrm{u}}\hat{\mathrm{u}}'Z_1 &= n^{-1}Z_1'\mathrm{u}\mathrm{u}'Z_1 - \left(n^{-1}Z_1'\mathrm{u}\mathrm{u}'X\right)(X'X)^{-1}X'Z_1 \\ &\quad - \left(n^{-1}Z_1'\,X\right)(n^{-1}X'X)^{-1}\left(n^{-1}X'\mathrm{u}\mathrm{u}'Z_1\right) \\ &\quad + \left(n^{-1}Z_1'\,X\right)(n^{-1}X'X)^{-1}\left(n^{-1}X'\mathrm{u}\mathrm{u}'X\right)(n^{-1}X'X)^{-1}\left(n^{-1}X'Z_1\right). \end{aligned}$$

Under the regularity conditions and the assumptions in White (1980), the probability limits of all these matrix products can be consistently estimated if in place of uu′, we use $\hat{\Omega}_0 = \text{diag}\{\hat{u}_i^2\}$. Reverting back, this means that

$$n^{-1}Z_1'\, M_x\, \hat{\Omega}_0\, M_x\, Z_1 \;\to_p\; \text{plim}\big[n^{-1}Z_1' M_x \text{uu}' M_x Z_1\big] = \text{plim}\big[n^{-1}Z_1' \hat{\text{u}}\hat{\text{u}}' Z_1\big].$$

Thus, we have obtained that a consistent estimator of the matrix Q is (now eliminating the redundant scaling factors)

$$\hat{Q} = X_1'Z\,(Z'Z)^{-1}\begin{bmatrix} n^{-1}Z_1\, M_x\, \hat{\Omega}_0\, M_x\, Z_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}(Z'Z)^{-1}\, Z'X_1 \to_p Q.$$

This can be compacted. Consider the matrix, suitably bracketed,

$$\widehat{X}_1' M_x \hat{\Omega}_0 M_x \widehat{X}_1 = X_1' P_z M_x \hat{\Omega}_0 M_x P_z X_1 = X_1' Z\,(Z'Z)^{-1}\big[Z'\, M_x \hat{\Omega}_0 M_x\, Z\big](Z'Z)^{-1} Z' X_1.$$

The outer terms are identical to the outer terms of $\hat{Q}$. Its middle term $Z'\, M_x \hat{\Omega}_0 M_x\, Z$ can be decomposed,

$$Z'\, M_x \hat{\Omega}_0 M_x\, Z = \begin{bmatrix} Z_1' \\ X_2' \end{bmatrix} M_x \hat{\Omega}_0 M_x \begin{bmatrix} Z_1 & X_2 \end{bmatrix} = \begin{bmatrix} Z_1' M_x \hat{\Omega}_0 \\ X_2' M_x \hat{\Omega}_0 \end{bmatrix} \begin{bmatrix} M_x Z_1 & M_x X_2 \end{bmatrix}.$$

However, $M_x X_2 = X_2' M_x = \mathbf{0}$ so

$$Z'\, M_x \hat{\Omega}_0 M_x\, Z = \begin{bmatrix} Z_1\, M_x\, \hat{\Omega}_0\, M_x\, Z_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

This is identical to the middle component of $\hat{Q}$; thus, we arrive at

$$\begin{aligned} \hat{Q} &= n^{-1} X_1' Z\,(Z'Z)^{-1}\,\big[Z'\, M_x \hat{\Omega}_0 M_x\, Z\big]\,(Z'Z)^{-1}\, Z' X_1 \\ &= n^{-1} X_1' P_z\, M_x \hat{\Omega}_0 M_x\, P_z X_1 \\ &= n^{-1} \widehat{X}_1' M_x \hat{\Omega}_0 M_x \widehat{X}_1. \end{aligned}$$

Appendix D

We present here the details of the Monte Carlo (MC) study whose results we report in the main text. The study was conducted using the software "gretl". For the random number generator, we have used the seed 1930021000.

Table A1 contains the random variables that we have used as building blocks.

Table A1. Building blocks of MC simulation.

Symbol	Distribtuion	Description
U_1	F(20, 15)	Snedecor's F-distr. with d.f. 20 (num.) and 15 (denom.)
U_3	P(1)	Poisson with mean equal to 1
U_4	$\chi^2(3)$	Chi-square with 3 d.f.
U_5	N(−1, 2)	Normal with mean equal to −1 and st.dev. 2
U_6	$t(6)$	Student's-t with 6 d.f.
U_7	U(−2, 2)	Continuous Uniform in (−2, 2)
U_8	U(0, 2)	Continuous Uniform in (0, 2)
U_9	$\text{U}_\text{d}(0, 2)$	Discrete Uniform in {0, 1, 2}

We have generated one "unobservable" that creates endogeneity and one that it does not, two regressors that become endogenous when the correlated unobservable is used in the data generation process, one exogenous, and three instruments. Table A2 contains the generating expressions.

Table A2. Variables in the MC simulation.

Symbol	Expression	Status
L_1	$0.7U_6 + U_7$	Latent correlated
L_2	U_7	Latent uncorrelated
X_{11}	$U_1 + U_3 + U_6$	Endogenous given L_1
X_{12}	$0.5U_3 + U_5 - 0.5U_6$	Endogenous given L_1
X_2	$U_1 + U_5$	Exogenous
Z_{11}	$\sqrt{U_3} - U_1$	Instrument
Z_{12}	$\lvert U_5 \rvert$	Instrument
Z_{13}	$U_3 - U_5$	Instrument

We note that, even though U_5 is a Normal random variable, the instrument $Z_{12} = |U_5|$ is relevant (correlated with the endogenous variables), because U_5 has a non-zero mean, and so its absolute value is a Folded Normal, and remains correlated with U_5 (in contrast, if U_5 was a zero-mean Normal and consequently Z_{12} a Half Normal, their covariance would be zero).

As regards the scenarios of homoskedasticity, random heteroskedasticity, and groupwise heteroskedasticity, the error term (including the unobservable variable) was generated as shown in Table A3 ($s = 1$ implies that the correlated L_1 latent variable was used).

Table A3. Error terms in the MC simulation.

Symbol	Expression	Model
$u^{[1,s]}$	$\mathrm{N}(0,2) + 3L_s,\ s = 1,\ 2$	Homoskedasticity
$u^{[2,s]}$	$\mathrm{N}(0,1+U_8) + 3L_s,\ s = 1,\ 2$	Random Heteroskedasticity
$u^{[3,s]}$	$\mathrm{N}(0,1+U_9) + 3L_s,\ s = 1,\ 2$	Groupwise Heteroskedasticity

For these three setups, the dependent variable was generated as

$$Y^{[t,s]} = 1 - 5X_2 + 2X_{11} + 1.5X_{12} + u^{[t,s]}, \quad t = 1,2,3, \quad s = 1,2.$$

So, for example, $Y^{[2,2]}$ is the situation where we have random heteroskedasticity and no endogeneity; thus, it was used to assess the empirical size of the test for this specific heteroskedastic scenario.

For the conditional heteroskedasticity scheme, we used a random-coefficients model, with mean values of the random coefficients equal to the specified coefficients above, together with the homoskedastic error term $u^{[1,s]}$, namely

$$Y^{[4,s]} = \mathrm{N}(1,0.2) - \mathrm{N}(5,1)\cdot X_2 + \mathrm{N}(2,0.4)\cdot X_{11} + \mathrm{N}(1.5,0.3)\cdot X_{12} + u^{[1,s]}, \quad s = 1,2.$$

Decomposed, this leads to

$$\begin{aligned} Y^{[4,s]} &= 1 - 5X_2 + 2X_{11} + 1.5X_{12} + u^{[4,s]}, \\ u^{[4,s]} &= \mathrm{N}(0,0.2) - \mathrm{N}(0,1)\cdot X_2 + \mathrm{N}(0,0.4)\cdot X_{11} + \mathrm{N}(0,0.3)\cdot X_{12} + u^{[1,s]}, \quad s = 1,2. \end{aligned}$$

Since all conditional heteroskedasticity factors are scaled by independent zero-mean Normals, no additional source of endogeneity is created.

The general matrix Hausman statistic is

$$\hat{q}_{het} \;=\; \hat{\mathrm{u}}'\,\widehat{X}_1\left[\widehat{X}_1'\,M_x\,\hat{\Omega}\,M_x\widehat{X}_1\right]^{-1}\widehat{X}_1'\,\hat{\mathrm{u}}.$$

The OLS regressions regressed $Y^{[t,s]}$ on a constant and $X = (X_2 : X_{11} : X_{12})$. $\hat{\mathrm{u}}$ are the OLS residuals used also in computing $\hat{\Omega}$. $M_x = I_n - P_x$, where P_x is the projection matrix of X. Its diagonal elements h_{ii} were used for the variants of $\hat{\Omega}$. $X_1 = (X_{11} : X_{12})$

and $\widehat{X}_1 = P_z X_1$, where P_z is the projection matrix of a constant and of $(X_2 : Z_1 : Z_2 : Z_3)$. For the homoskedastic variant of the statistic, the estimated OLS error variance $\hat{\sigma}_u$ was used, instead of $\hat{\Omega}$.

Notes

[1] Sometimes it is also called the "artificial regression" or "control function" approach.

[2] In the literature, the test is presented with the use of the Moore–Penrose pseudo-inverse V^+, most likely because its uniqueness avoids the necessity to choose among alternatives in an *ad hoc* manner, as well as the uncertainty of obtaining possibly different results for different generalized inverses in finite samples. Regardless, the limiting distributional result holds for any generalized inverse, see Hausman and Taylor (1981).

[3] The need for a generalized inverse in the original formulation of the test is treated as "cumbersome" in the literature, see for example Greene (2012, p. 276) and Wooldridge (2002, p. 119), and it is also put forth as an argument to favor the use of the augmented regression test.

[4] The "augmented regression" test also guards against this possibility, since it uses the residuals from regressing each endogenous variable on the instruments. If exact linear dependence exists, the related series of residuals will be a series of zeros.

[5] This monotonic fall of power, as we "intensify" the degree to which we attempt to correct the heteroskedasticity estimator for finite sample performance, is in accord with what MacKinnon (2013, pp. 456–57) found.

[6] In case there is an issue with the validity of the instruments, as discussed earlier, in the augmented regression method, we would get at least one series of zero residuals.

[7] So, as regards the heteroskedasticity corrector HC1, the number of regressors in the augmented regression setup is $k = 2K_1 + K_2$, while for HC2 and HC3, the diagonal element h_{ii} is of a projection matrix that includes these additional variables.

[8] MacKinnon (2013, pp. 449–52) also found in his simulations that the HC3 variant performs best as regards empirical size in small samples.

References

Adkins, Lee C., Randall C. Campbell, Viera Chmelarova, and R. Carter Hill. 2012. The Hausman test, and some alternatives, with heteroskedastic data. In *Essays in Honor of Jerry Hausman*. Advances in Econometrics, vol. 29. Leeds: Emerald Group Publishing Ltd.

Amini, Shahram, Michael S. Delgado, Daniel J. Henderson, and Christopher F. Parmeter. 2012. Fixed vs. random: The Hausman test four decades later. In *Essays in Honor of Jerry Hausman*. Advances in Econometrics, vol. 29. Leeds: Emerald Group Publishing Ltd.

Durbin, James. 1954. Errors in variables. *Revue de l'institut International de Statistique* 22: 23–32. [CrossRef]

Greene, William H. 2012. *Econometric Analysis*, 7th ed. Harlow: Pearson Education Ltd.

Hahn, Jinyong, John C. Ham, and Hyungsik Roger Moon. 2011. The Hausman test and weak instruments. *Journal of Econometrics* 160: 289–99. [CrossRef]

Hausman, Jerry A. 1978. Specification tests in econometrics. *Econometrica* 46: 1251–71. [CrossRef]

Hausman, Jerry A., and William E. Taylor. 1981. A generalized specification test. *Economics Letters* 8: 239–45. [CrossRef]

MacKinnon, James G. 2013. Thirty years of heteroskedasticity-robust inference. In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*. Edited by Xiaohong Chen and Norman R. Swanson. New York: Springer, pp. 437–61.

White, Halbert. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–38. [CrossRef]

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of cross Section and Panel Data*. Cambridge: MIT Press.

Wu, De-Min. 1973. Alternative tests of independence between stochastic regressors and disturbances. *Econometrica* 41: 733–50. [CrossRef]

Wu, De-Min. 1974. Alternative tests of independence between stochastic regressors and disturbances: Finite sample results. *Econometrica* 42: 529–46. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.