Advice on using heteroscedasticity based identification

Christopher F. Baum and Arthur Lewbel
Boston College

February 12, 2019

Abstract

Lewbel (2012) provides a heteroscedasticity based estimator for linear regression models containing an endogenous regressor when no external instruments or other such information is available. The estimator is implemented in the Stata module ivreg2h by Baum and Schaffer (2012). This note gives some advice and instructions to researchers who want to use this estimator.

1 Introduction

Linear regression models containing endogenous regressors are generally identified using outside information such as exogenous instruments, or by parametric distribution assumptions. Some papers obtain identification without external instruments by exploiting heteroscedasticity, including Rigobon (2003), Klein and Vella (2010), Lewbel (1997, 2018) and Prono (2014). In particular, Lewbel (2012) shows how one can use heteroskedasticity to construct instruments when no external instruments are available. Other papers that obtain identification using constructed instruments include Lewbel (1997) and Erickson and Whited (2002). See Lewbel (in press) for a general discussion of identification methods like these.

In this note, we provide advice and instructions for researchers who wish to apply the Lewbel (2012) estimator. That article includes estimators for fully simultaneous systems, semiparametric systems, and bounds for when

key identifying assumptions are violated. However, most empirical applications use the estimator for a single-equation linear regression model with a single endogenous regressor, which is the focus here. This linear single equation estimator has been implemented by Baum and Schaffer (2012) as the Stata module `ivreg2h`, which is available from the SSC Archive.

2 The model and estimator

Assume a sample of observations of endogenous variables Y_1 and Y_2 and a vector of exogenous covariates X. We wish to estimate γ and the vector β in the model

$$\begin{aligned} Y_1 &= X'\beta + Y_2\gamma + \varepsilon_1 \\ Y_2 &= X'\alpha + \varepsilon_2 \end{aligned}$$

where the errors ε_1 and ε_2 may be correlated.

Standard instrumental variables estimation depends on having an element of X that appears in the Y_2 equation but not in the Y_1 equation, and uses that excluded regressor as an instrument for Y_2. The problem considered here is that perhaps no element of X is excluded from the Y_1 equation, or equivalently, we're not sure that any element of β is zero. Lewbel (2012) provides identification and a corresponding very simple linear two stage least squares estimator for β and γ in this case where no element of X can be used as an excluded instrument for Y_2. The method consists of constructing valid instruments for Y_2 by exploiting information contained in heteroscedasticity of ε_2.

In addition to the standard exogenous X assumptions that $E(X\varepsilon_1) = 0$, $E(X\varepsilon_2) = 0$, and $E(XX')$ is nonsingular, the key additional assumptions required for applying the Lewbel (2012) estimator are that $Cov(Z, \varepsilon_1\varepsilon_2) = 0$ and $Cov(Z, \varepsilon_2^2) \neq 0$, where either $Z = X$ or Z is a subset of the elements of X.

The Lewbel (2012) estimator can be summarized as the following two steps.

1. Estimate $\widehat{\alpha}$ by an ordinary least squares regression of Y_2 on X, and obtain estimated residuals $\widehat{\varepsilon}_2 = Y_2 - X'\widehat{\alpha}$.

2. Let Z be some or all of the elements of X (not including the constant term). Estimate β and γ by an ordinary linear two stage least squares

regression of Y_1 on X and Y_2, using X and $(Z - \overline{Z})\widehat{\varepsilon}_2$ as instruments, where $\overline{Z}$ is the sample mean of Z.

This estimator is implemented in the Stata module `ivreg2h` by Baum and Schaffer (2012). Note that applying the estimator requires choosing which elements of X will comprise the vector Z used to construct instruments. The default assumption in `ivreg2h` is that Z includes all of the elements of X except for the constant term. However, one might also choose to let Z be only some of the elements of X, if doing so helps to satisfy the assumptions required for the estimator, as discussed in the next section.

3 Advice on applying the estimator

The main question to be answered by applied researchers who wish to use this estimator is whether the key assumptions, that $Cov(Z, \varepsilon_1\varepsilon_2) = 0$ and $Cov(Z, \varepsilon_2^2) \neq 0$, are likely to hold. Below we discuss conditions that are sufficient to make these key assumptions hold. The virtue of these sufficient conditions (given as Assumptions A1, A2, and A3 below) is that each can either be motivated by economic theory, or can be empirically tested with data, or both. It is possible for the key assumptions to hold without satisfying Assumptions A1, A2, and A3. However, if you can provide evidence (theory and tests as we describe below) for why these sufficient conditions should hold in your application, then the estimator is more likely to be appropriate for you to use.

ASSUMPTION A1: The errors ε_1 and ε_2 have the factor structure

$$\begin{aligned} \varepsilon_1 &= cU + V_1 \\ \varepsilon_2 &= U + V_2 \end{aligned}$$

where c is a constant, and U, V_1, and V_2 are unobserved error terms that are mutually independent conditional on Z.

The interpretation of Assumption A1 is that Y_2 is endogenous because it contains an error component U that appears in the errors of both equations. This assumption is not directly testable, and so must be justified by economic or econometric theory. To illustrate, here we provide examples of how Assumption A1 could be justified in a variety of contexts.

Example: Suppose Y_2 is endogenous because it is mismeasured. Then V_1 is the true outcome model error, and U is the measurement error. Classical measurement error in linear regression models satisfies Assumption A1.

Example: Suppose Y_1 is a wage, and Y_2 is education level. Here U could be unobserved ability, which affects both educational attainment Y_2 and one's wage Y_1. Then V_1 represents all the unobservables that affect wages but not education, while V_2 represents all the unobservables that affect education but not wages.

Example: Suppose Y_1 is a firm's value added output per unit of capital, and Y_2 is the firm's labor per unit of capital. Here U could be unobserved entrepreneurship, which affects both productivity and the chosen level of inputs. Then V_1 represents all the unobservables that affect productivity but not inputs, and vice versa for V_2.

The point here, as illustrated by these examples, is that the endogeneity of Y_2 takes the form of there being some underlying, unobserved factor U that affects both Y_1 and Y_2.

ASSUMPTION A2: U^2 is not correlated with Z.

Assumption A2 says that U is homoscedastic. The Y_1 equation is a structural model, so if we can argue that it is correctly specified without important omitted variables, then it is common to assume remaining errors are completely idiosyncratic. This may be a difficult assumption to justify in theory, but it is partly testable, in particular, we may apply a Pagan and Hall (1983) test to the Y_1 equation. The commonly used Breusch and Pagan (1979) and White (1980) tests are not appropriate in the context of the Y_1 equation, as it contains an endogenous regressor. See Baum, Schaffer, and Stillman (2003), Section 3, for more detail. The form of the Pagan–Hall test that allows specification of the included regressors should be used, as just the variables in Z should be included in the test of this assumption that (U^2 is not correlated with Z). Having heteroscedasticity of a form where U^2 is correlated with regressors other than Z would not violate Assumption A2.

A limitation of this test is that it tests homoscedasticity of ε_1, so if we reject homoscedasticity, we can't know if the rejection is due to violating Assumption A2 or whether it is due to harmless heteroscedasticity of V_1. In short, failing to reject homoscedasticity of ε_1 provides evidence supporting Assumption A2, but rejecting homoscedasticity of ε_1 does not mean that Assumption A2 is necessarily violated.

Note that Assumption A2 does not require that U^2 be fully homoscedas-

tic: only that it is not correlated with Z. As discussed at the end of the previous section, to satisfy Assumption A2 (and A3 below) one might be selective about which elements of X to include in Z.

ASSUMPTION A3: ε_2^2 is correlated with Z.

This assumption is needed to ensure that the constructed instrument ends up correlated with Y_2. If the previous assumptions hold, then this assumption is equivalent to heteroscedasticity of V_2 relative to Z. This assumption is easy to justify, since the Y_2 equation need not be a structural equation. The Y_2 equation is like the first stage of two stage least squares: it can be defined as just a linear projection of Y_2 on exogenous covariates. Moreover, this assumption can be tested by applying a Breusch and Pagan (1979) test to the Y_2 equation.[1] Unlike the test of Assumption A2 for the Y_1 equation, to satisfy Assumption A3 we want to reject homoscedasticity.[2]

Note that the above assumptions are not strictly necessary for the estimator, e.g., it is possible that the factor model of Assumption A1 does not hold but the estimator is still consistent (see Lewbel (2018) for an example). However, we can have more confidence that the estimator is consistent in a given application if we can argue that the logic of Assumption A1 holds and if we pass the tests in Assumptions A2 and A3.

Additional tests, lending even more support for the estimator, are possible when Z has more than one element. In that case, the model is overidentified, and one can then apply standard overidentification tests such as the Hansen (1982) and Sargan (1958) J-test. However, it is important to note that this only tests a necessary condition for validity of the method, which is that all instruments yield the same coefficient estimates. It is possible, e.g., that one fails to reject overidentification tests not because the assumptions hold, but because the constructed instruments happen to all yield the same incorrect coefficient estimates. Still, failing to reject overidentification tests provides additional evidence in support of the model and estimator.

To summarize the results of this section, one way to use this estimator convincingly is to:

[1] As there are no endogenous regressors in the Y_2 equation, the standard heteroscedasticity tests may be used. The Pagan and Hall (1983) test could also be used, as it is equivalent to the Breusch–Pagan test when applied to an OLS equation.

[2] The Breusch and Pagan (1979) test is preferred over the general White (1980) test, as it allows us to target the necessary form of heteroscedasticity, i.e., correlation of the squared error with Z.

1. Use economic theory and/or data to justify linearity of the model $Y_1 = X'\beta + Y_2\gamma + \varepsilon_1$ and the assumption that X is exogenous.

2. Use economic theory and/or data to justify the factor structure of the errors given by Assumption A1.

3. Choose a set of covariates Z (either all the elements of X except the constant, or some subset of those elements) to use for constructing the instruments $(Z - \overline{Z})\widehat{\varepsilon}_2$. For the chosen Z, apply theory and the above described tests to justify the remaining identifying assumptions.

4 Implementing the estimator and tests

Using the Lewbel (2012) method, instruments are constructed as simple functions of the model's data. This approach may be (a) applied when no ordinary (external) instruments are available, or, alternatively, (b) used along with external instruments to improve the efficiency of the instrumental variables estimator. Constructed instruments along with external instruments can also be used to obtain overidentification, thereby allowing application of Sargan–Hansen tests (of the orthogonality conditions or overidentifying restrictions) which would not be possible in the case of exact identification by external instruments. This then allows one to simultaneously test validity of both the external instruments and the constructed instruments.

The implementation of the estimator in `ivreg2h` is based on the earlier `xtivreg2` (Schaffer (2015)) and `ivreg2` (Baum, Schaffer, and Stillman (2003, 2007)) routines. Essentially, `ivreg2h` generates the heteroscedasticity based constructed instruments, and then implements instrumental variables estimation like these earlier routines. In addition to pure cross section or time series data, `ivreg2h` can also be applied to panel data using the within transformation of a fixed effects model: see the `fe` option described below. As `ivreg2h` is a variant of `ivreg2`, essentially all of the features and options of that program are available in `ivreg2h`. For that reason, you can consult `help ivreg2` for full details of the available options.

The `robust` and `gmm2s` options should generally be employed, invoking the IV-GMM estimator. This will compute the Hansen J statistic as a test of overidentifying restrictions. The default Sargan test assumes normality of the errors. See Baum, Schaffer, and Stillman (2003, 2007) for further details.

Note that the gmm2s option supersedes the gmm option described in the earlier article.

The ivreg2h program provides four additional options: gen, gen(*string[,replace]*), fe and z(). If the gen option is given, the generated (constructed) instruments are saved, with names built from the original variable names suffixed with _g. If greater control over the naming of the generated instruments is desired, use the gen(*string[,replace]*) option. The string argument allows the specification of a stub, or prefix, for the generated variable names, which will also be suffixed with _g. You can remove earlier instruments with those same names with the replace suboption. If the data have been declared as a panel, you can use the fe option to specify that a fixed-effects model should be estimated, as in xtivreg2. The z() option can be used to specify that only some of the included exogenous variables should be used to generate instruments, as suggested above.

The ivreg2h program can be invoked to estimate either (a) a model that would be identified even without the constructed instruments, or (b) a model that, without constructed instruments, would fail the order condition for identification by either having no excluded instruments, or by having fewer excluded instruments than needed for traditional identification.

In case (a), where an adequate number of external instruments are augmented by the generated constructed instruments, the program provides three sets of estimates: the traditional IV estimates, the estimates using only the generated instruments, and the estimates using both generated and excluded instruments. In this case, ivreg2h automatically produces a Hayashi C test of the excluded instruments' validity, equivalent to that provided by the orthog() option in ivreg2).[3] The results of the third estimation (the one including both generated and excluded instruments) are saved in the ereturn list. All three sets of estimates are saved, named StdIV, GenInst and GenExtInst, respectively.

In case (b), where the equation would be underidentified without constructed instruments, either one or two sets of estimates will be produced and displayed. If there are no excluded instruments, only the estimates using the generated instruments are displayed. If there are excluded instruments but too few to produce identification by the order condition, the estimates using only generated instruments and those produced by both generated and excluded instruments will be displayed. Unlike ivreg2 or ivregress, ivreg2h

[3]Baum, Schaffer, and Stillman (2003),18–19.

allows the syntax

```
ivreg2h depvar exogvar (endogvar=) [if exp] [in range],  options
```

as after augmentation with the generated regressors, the order condition for identification will be satisfied. The resulting estimates are saved in the `ereturn list` and as a set of estimates named `GenInst` and, optionally, `GenExtInst`.

The Pagan and Hall (1983) tests referenced above are available from the `ivreg2` package of Baum, Schaffer, and Stillman (2003) using the `ivhettest` command. The default test does not assume normality of the errors.

4.1 Saved results

In the `estimates table` output, the displayed results `j, jdf` and `jp` refer to the Hansen J statistic, its degrees of freedom, and its p-value. If i.i.d. errors are assumed and a Sargan test is displayed in the standard output, the Sargan statistic, its degrees of freedom and p-value are displayed in `j, jdf` and `jpval`, as the Hansen and Sargan statistics coincide in that case. The results of the most recent estimation are saved in the `ereturn list`.

5 Examples of usage

In this example from Lewbel (2012), centering of regressors is only used to match the published results.

```
ssc install center // (if needed)
ssc install bcuse // (if needed)
bcuse engeldat
center age-twocars, prefix(z_)
ivreg2h foodshare z_* (lrtotexp=), small robust
ivreg2h foodshare z_* (lrtotexp = lrinc), small robust
ivreg2h foodshare z_* (lrtotexp = lrinc), small robust gmm2s z(z_age-z_age2sp)
```

Example of use with panel data and HAC standard errors:

```
webuse grunfeld, clear
ivreg2h invest L(1/2).kstock (mvalue=), fe
ivreg2h invest L(1/2).kstock (mvalue=L(1/4).mvalue), fe robust bw(2)
```

6 Additional comments

Here we provide answers to additional questions that have been asked about the estimator.

1. Can validity of the estimator be tested?

   Partially. The tests discussed in the previous sections are examples.

2. What if Y_1 or Y_2 is discrete?

   It is possible that the estimator will still be valid in this case. Lewbel (2018) gives one set of conditions that suffice for validity of the estimator. However, the factor structure given by Assumption A1 will generally not hold if Y_1 or Y_2 is discrete, so it is much harder to justify application of the estimator. One might still apply the tests discussed in the previous section to provide some evidence to rationalize the estimator in this case.

3. What does it mean if coefficient estimates are close to those from ordinary least squares?

   In any application of instrumental variables estimators, coefficient estimates can be close to ordinary least squares either by chance, or if the instruments are highly correlated with the endogenous regressors. The same is true of constructed instruments.

4. Can the estimator be used with more than one endogenous regressor?

   Conditions for validity of the estimator have been proven for one endogenous regressor. The estimator may be valid with multiple endogenous regressors, but the exact conditions required for validity in that case have not been shown.

5. What if I have additional instruments?

   This is the best case scenario, as those external instruments can be used along with the constructed instruments in the second step of the estimator (as discussed earlier). In particular, one of the best uses of the constructed instruments is to provide overidentifying information for model tests and for robustness checks. For example, one could apply the overidentification tests discussed in the previous sections to estimates based on both constructed and external instruments. If validity is rejected, then either the model is misspecified or at least one of these instruments is invalid. If validity is not rejected, it's still possible that the model is wrong or the instruments are invalid, but one would at least have increased confidence that both the external instruments and the constructed instruments are valid. More informally, one

might simply compare the estimated coefficients based on constructed instruments versus those based on external instruments.[4] If they are numerically similar, that increases confidence in the robustness of the model, as the two estimators based on very different identifying assumptions are yielding similar results. More generally, identification based on constructed instruments is preferably not used in isolation, but rather is ideally employed in conjunction with other means of obtaining identification, both as a way to check robustness of results to alternative identifying assumptions and to increase the efficiency of estimation.

7 Conclusions

In the few years since the heteroskedasticity-based estimator was proposed, it has been cited more than five hundred times according to Google Scholar. But like any identification method that is based largely on structure and functional form, one must be very cautious about interpreting the results. This note should help ensure that the estimator is applied appropriately.

References

Baum, C. F. and M. E. Schaffer. 2012. IVREG2H: Stata module to perform instrumental variables estimation using heteroskedasticity-based instruments. Technical Report Statistical Software Components S457555, Boston College.

Baum, C. F., M. E. Schaffer, and S. Stillman. 2003. Instrumental variables and GMM: Estimation and testing. *Stata Journal* 3(1): 1–31.

—. 2007. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *Stata Journal* 7(4): 465–506.

Breusch, T. S. and A. R. Pagan. 1979. A simple test for heteroskedasticity and random coefficient variation. *Econometrica* 47: 1287–1294.

[4] As discussed earlier, these alternative estimates are automatically provided by `ivreg2h`.

Erickson, T. and T. M. Whited. 2002. Two-step GMM estimation of the errors-in-variables model using high-order moments. *Econometric Theory* 18(3): 776–799.

Hansen, L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50(3): 1029–1054.

Klein, R. and F. Vella. 2010. Estimating a class of triangular simultaneous equations models without exclusion restrictions. *Journal of Econometrics* 154(42): 154–164.

Lewbel, A. 1997. Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D. *Econometrica* 65: 1201–1213.

—. 2012. Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business and Economic Statistics* 30: 67–80.

—. 2018. Identification and Estimation Using Heteroscedasticity Without Instruments: The Binary Endogenous Regressor Case. *Economics Letters* 165: 10–12.

—. in press. The Identification Zoo – Meanings of Identification in Econometrics. *Journal of Economic Literature* .

Pagan, A. R. and D. Hall. 1983. Diagnostic tests as residual analysis. *Econometric Reviews* 2(2): 159–218.

Prono, T. 2014. The Role of Conditional Heteroskedasticity in Identifying and Estimating Linear Triangular Systems, with Applications to Asset Pricing Models that Include a Mismeasured Factor. *Journal of Applied Econometrics* 29(5): 800–824.

Rigobon, R. 2003. Identification Through Heteroskedasticity. *Review of Economics and Statistics* 85(4): 777–792.

Sargan, J. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26(3): 393–415.

Schaffer, M. E. 2015. XTIVREG2: Stata module to perform extended IV/2SLS, GMM and AC/HAC, LIML and k-class regression for panel data models. Technical Report Statistical Software Components S456501, Boston College.

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–838.