Advice on using heteroscedasticity based identi cation

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 Scha er (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 identi ed using outside information such as exogenous instruments, or by parametric distribution assumptions. Some papers obtain identi cation 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 identi cation using constructed instruments include Lewbel (1997) and Erickson and Whited (2002). See Lewbel (in press) for a general discussion of identi cation 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 Scha er (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 and the vector in the model

$$
\begin{aligned}
Y_1 &= X^0 + Y_2 + "_1 \\
Y_2 &= X^0 + "_2
\end{aligned}
$$

where the errors $"_1$ and $"_2$ 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 is zero. Lewbel (2012) provides identi cation and a corresponding very simple linear two stage least squares estimator for and 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 $"_2$.

In addition to the standard exogenous X assumptions that $E(X"_1) = 0$, $E(X"_2) = 0$, and $E(XX^0)$ is nonsingular, the key additional assumptions required for applying the Lewbel (2012) estimator are that $Cov(Z; "_1"_2) = 0$ and $Cov(Z; "_2^2) \neq 0$, 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 $\hat{}$ by an ordinary least squares regression of Y_2 on X, and obtain estimated residuals $\hat{"}_2 = Y_2 - X^0\hat{}$.

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

regression of Y_1 on X and Y_2, using X and (Z

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

1. Use economic theory and/or data to justify linearity of the model $Y_1 = X'\beta_1 + Y_2\gamma_1 + \varepsilon_1$ 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 $(Z - \bar{Z})\hat{\varepsilon}_2$. 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 or-

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 su ce 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 coe cient estimates are close to those from ordinary least squares?

In any application of instrumental variables estimators, coe cient 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 overidenti cation tests discussed in the previous sections to estimates based on both constructed and external instruments. If validity is rejected, then either the model is misspeci ed 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 con dence 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. Identi cation and Estimation Using Heteroscedasticity Without Instruments: The Binary Endogenous Regressor Case. Economics Letters 165: 10{12.

|. in press. The Identi cation Zoo { Meanings of Identi cation 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. Identi cation 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.

Scha er, 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.