Identification and Estimation Using Heteroscedasticity Without Instruments: The Binary Endogenous Regressor Case

Arthur Lewbel*
Boston College

December 2016

Abstract

Lewbel (2012) provides an estimator for linear regression models containing an endogenous regressor, when no outside instruments or other such information is available. The method works by exploiting model heteroscedasticity to construct instruments using the available regressors. Some authors have considered the method in empirical applications where an endogenous regressor is binary (e.g., endogenous Diff-in-Diff or endogenous binary treatment models), without proving validity of the estimator in that case. The present paper shows that the assumptions required for Lewbel's estimator can indeed be satisfied when an endogenous regressor is binary.

JEP codes: C35, C36, C30, C13

Keywords: Simultaneous systems, linear regressions, endogeneity, identification, heteroscedasticity, binary regressors, dummy regressors, linear probability model, logit, probit.

*Corresponding address: Arthur Lewbel, Dept of Economics, Maloney Hall, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu, https://www2.bc.edu/arthur-lewbel/

1 Introduction

Linear regression models containing endogenous regressors are generally identified and estimated using either outside information, such as exogenous instruments, or by parametric distribution assumptions (known as identification by functional form). A few papers obtain identification without outside instruments by exploiting heteroscedasticity, including Rigobon (2003), Klein and Vella (2010), Lewbel (2012), and Prono (2014). Other methods of obtaining identification, such as taking advantage of higher moments or potential nonlinearities, are discussed in Lewbel's (2016) survey article.

Some authors have questioned whether these methods can be applied to the case where the endogenous regressor is binary. For example, Diff-in-Diff models contain binary regressors by construction, and are only valid without instrumenting given exogeneity assumptions. Similarly, endogenous treatment indicators are often binary.

Examples of authors who questioned whether the Lewbel (2012) estimator can be used with binary endogenous variables include Emran, Robano, and Smith (2014) and Hoang, Pham, and Ulubaşoğlu (2014). Others have applied the Lewbel (2012) estimator with a binary endogenous variable, though without being able to verify if all the assumptions hold. See, e.g., Le Moglie, Mencarini, and Rapallini, (2015).

This paper shows that the Lewbel (2012) estimator may sometimes be used in such cases, by providing one set of conditions that suffice for validity of the estimator when an endogenous regressor is binary. So, e.g., the estimator might be applied to estimate a (homogeneous) treatment effect when the binary treatment is not randomly assigned and when exogenous instruments are not available. The advantage of this result is that it shows that the assumptions of the Lewbel (2012) estimator are not necessarily violated when an endogenous regressor is binary. The disadvantage is that, unlike the case with a continuous endogenous regressor, the sufficient conditions given here for a binary endogenous regressor impose a very strong distribution restriction on the error term of the regression equation. Of course, this result just shows existence, not uniqueness, of assumptions that work. It does not rule out the possibility that the estimator's assumptions could also be satisfied by alternative, less restrictive assumptions in the presence of a binary endogenous regressor.

2 The Model

Suppose we observe a sample of observations of endogenous variables Y_1 and Y_2, and a vector of exogenous covariates X. We wish to estimate the parameter γ and the parameter 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. As in Lewbel (2012), we will also consider the more general case where $Y_2 = g(X) + \varepsilon_2$ for some nonlinear, possibly unknown function g.

The standard instrumental variables solution to estimating β and γ is to find an element of X that appears in the Y_2 equation but not in the Y_1 equation, and use that excluded regressor as an instrument for Y_2. The problem for identification and estimation considered here is that 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 the case where no element of X is excluded from the Y_1 equation, so no element of X can be used as an instrument for Y_2. The method consists of constructing valid instruments for Y_2 by exploiting information contained in heteroscedasticity of ε_2.

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. 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.

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 this 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. Lewbel (2012) shows that a variety of standard econometric models satisfy these assumptions. For example, the assumptions hold

when the errors ε_1 and ε_2 satisfy the factor structure $\varepsilon_1 = cU + V_1$, and $\varepsilon_2 = U + V_2$ for some constant c, where U and V_1 are unobserved homoscedastic errors, V_2 is an unobserved heteroscedastic error, and U, V_1, and V_2 are mutually independent conditional on Z. Examples where these conditions hold are when Y_2 is endogenous due to classical measurement error, or because of the presence of some underlying unobservable factor U that affects both Y_1 and Y_2 (e.g., U could be unobserved ability in a model where Y_1 is education and Y_2 is a labor market outcome).

Lewbel (2012) doesn't explicitly assume that Y_2 is continuous. However, that paper doesn't show that its identifying assumptions can be satisfied when Y_2 is not continuous. For example, if Y_2 was binary then U could not be independent of V_2 in the above factor structure example. What the next section shows is that the identifying assumptions can be satisfied when Y_2 is binary.

3 A Binary Endogenous Regressor

Suppose that Y_2 is binary. Then $Y_2 = X'\alpha + \varepsilon_2$ is a linear probability model. But we also wish to allow for more general models, so let $Y_2 = g(X) + \varepsilon_2$ where $g(X) = E(Y_2 \mid X)$. Here $g(X)$ is some possibly nonlinear and possibly unknown function. For example, if Y_2 satisfies a probit or logit model, then $g(X) = F\left(X'\alpha\right)$ where F is the cumulative normal or logistic distribution function. Also included are nonparametric models, where $g(X)$ is estimated by a nonparametric regression of Y_2 on X. Note in particular that Y_2 could be an indicator of treatment that might not be randomly assigned. In that case estimation of γ corresponds to estimation of a (homogeneous) treatment effect.

Regardless of whether we estimate a linear probability model regression where the estimate is $\widehat{g}(X) = X'\widehat{\alpha}$, or let $\widehat{g}(X) = F\left(X'\widehat{\alpha}\right)$ where $\widehat{\alpha}$ is obtained by a logit, probit, or other threshold crossing model estimator, or estimate $\widehat{g}(X)$ nonparametrically by, e.g., a kernel or sieve estimator, once we obtain estimates of the residuals $\widehat{\varepsilon}_2 = Y_2 - \widehat{g}(X)$, step 2 of the estimator described above remains the same.

We maintain the usual linear model assumptions for the exogenous regressors X, i.e., that X is uncorrelated with ε_1 and ε_2, and that $E(XX')$ is nonsingular. If g is nonlinear (as in a logit, probit, or nonparametric regression model) then assume whatever is needed for consistent estimation of g. We now

show how the key additional assumptions required for the Lewbel (2012) estimator can be satisfied with Y_2 binary. For simplicity, the result is derived taking $Z = X$, which then implies that the restrictions can also hold when Z is any subset of X.

ASSUMPTION A1: Let $g(X) = E(Y_2 \mid X)$ and define $\varepsilon_2 = Y_2 - g(X)$. Assume $g(X)$ is finite and that $Cov\left[X, g(X)(1 - g(X))\right] \neq 0$.

ASSUMPTION A2: Assume $Y_1 = X'\beta + Y_2\gamma + \varepsilon_1$ with $\varepsilon_1 = Y_2 U + V$ for some unobserved random errors U and V, where Y_2, U, and V are conditionally mutually independent, conditioning on X. Assume $E(U \mid X) = c(X) / (g(X)(1 - g(X)))$ and $E(V \mid X) = -c(X) / (1 - g(X))$, where $c(X)$ is any function such that $Cov(X, c(X)) = 0$.

Assumption A1 imposes minimal restrictions on Y_2 and X, and hence on the error ε_2. The covariance condition in Assumption A1 is testable, since it can be estimated as the sample covariance between X and $\widehat{g}(X)(1 - \widehat{g}(X))$. In contrast, Assumption A2 places strong distributional restrictions on ε_1, specifically, on the conditional means of the component latent errors U and V. Again it should be stressed that these are not necessary conditions. Rather, they're just one possible set of assumptions that can be shown to work.

Note that the covariance condition in Assumption A2 will automatically hold if $c(X)$ is constant. However, it's also easy to find functions $c(X)$ that can work. For example if Z is any symmetrically distributed element of X that is independent of the other elements of X, then $c(X)$ could equal $(Z - E(Z))^k$ for any even integer k.

THEOREM 1: Let Assumptions A1 and A2 hold. Then $E(\varepsilon_1 \mid X) = 0$, $E(\varepsilon_2 \mid X) = 0$, $Cov(X, \varepsilon_1\varepsilon_2) = 0$ and $Cov(X, \varepsilon_2^2) \neq 0$.

PROOF: Verifying each of the conditions in turn, we have

$$\begin{aligned} E(\varepsilon_1 \mid X) &= E(Y_2 U + V \mid X) = g(X) E(U \mid X) + E(V \mid X) \\ &= g(X) \frac{c(X)}{g(X)(1 - g(X))} + \frac{-c(X)}{(1 - g(X))} = 0. \end{aligned}$$

$$E(\varepsilon_2 \mid X) = E(Y_2 - g(X) \mid X) = g(X) - g(X) = 0.$$

$$\begin{aligned} E(\varepsilon_1\varepsilon_2 \mid X) &= E(Y_2 U\varepsilon_2 + V\varepsilon_2 \mid X) = E(Y_2 U\varepsilon_2 \mid X) + E(V \mid X) E(\varepsilon_2 \mid X) \\ &= E(Y_2 U(Y_2 - g(X)) \mid X) = E(U(Y_2 - Y_2 g(X)) \mid X) \\ &= E(U \mid X) g(X)(1 - g(X)) = \frac{c(X)}{g(X)(1 - g(X))} g(X)(1 - g(X)) = c(X) \end{aligned}$$

so

$$Cov(X, \varepsilon_1\varepsilon_2) = Cov(X, E(\varepsilon_1\varepsilon_2 \mid X)) = Cov(X, c(X)) = 0.$$

and

$$\begin{aligned} E\left(\varepsilon_2^2 \mid X\right) &= E\left((Y_2 - g(X))^2 \mid X\right) = E\left(\left(Y_2 - 2Y_2 g(X) + g(X)^2\right) \mid X\right) \\ &= g(X) - 2g(X)^2 + g(X)^2 = g(X)(1 - g(X)) \end{aligned}$$

so

$$Cov(X, \varepsilon_2^2) = Cov\left(X, E\left(\varepsilon_2^2 \mid X\right)\right) = Cov\left[X, g(X)(1 - g(X))\right] \neq 0.$$

Using the same types of derivations as in the above proof, one can also readily verify that $E(\varepsilon_1 Y_2) = E(c(X))$ so Y_2 is indeed an endogenous regressor as long as $E(c(X)) \neq 0$. Theorem 1, along with the results in Lewbel (2012), establishes that the constructed instruments $(X - \overline{X})\widehat{\varepsilon}_2$, are valid, and so the estimator of regressing of Y_1 on X and Y_2, using X and $(X - \overline{X})\widehat{\varepsilon}_2$ as instruments (or any subset of the elements of $(X - \overline{X})\widehat{\varepsilon}_2$ as instruments) can be applied.

4 Extensions and Conclusions

A common problem in empirical work is the presence of binary regressors that might be endogenous. For example, Diff-in-Diff models contain binary regressors by construction, and are only valid without instrumenting given exogeneity assumptions. Similarly, endogenous treatment indicators are often binary.

When instruments are not available, alternative methods for obtaining identification may be desired. Theorem 1 in this paper shows that the assumptions required for the estimator in Lewbel (2012) can be satisfied in the presence of a binary endogenous regressor Y_2.

A drawback of this result is that, unlike the case for continuous Y_2, there is no obvious behavioral model that would imply that Assumption A2 and hence Theorem 1 holds. When Y_2 is binary, the estimator might not perform well under more standard assumptions. However, Theorem 1 is only meant to show that it's possible for the estimator to work, i.e., that the required assumptions can be satisfied. There may well exist more plausible or better motivated alternative constructions to Theorem 1 that would also work. Searching for such alternatives would be a useful direction for future research. To apply the estimator, one must only assume that the general conditions given in Lewbel (2012) hold, not that the specific necessary conditions used for Theorem 1 are satisfied. Given these assumptions, existing implementations of the estimator, such as the STATA module IVREG2H, can be applied with a binary Y_2.

Theorem 1 assumed a binary Y_2, but similar constructions would be possible for alternative distributions, such as when Y_2 is discrete with more support points, or is a censored variable, or more generally may be both continuous over some regions and contain mass points.

Theorem 1 applies regardless of the specification of the function $g(X)$, and so in particular can be used with the linear probability model $Y_2 = X'\alpha + \varepsilon_2$, if one is willing to assume that $E(Y_2 \mid X) = X'\alpha$. See, however, Lewbel, Dong, and Yang (2012) for warnings regarding the linear probability model. When $g(X)$ is nonlinear (e.g., if it's given by a logit or probit model), then alternative estimators exist. For example, in that case one could use $\widehat{g}(X)$, an estimate of $g(X)$, as an instrument for Y_2. See, e.g., Dong (2012) and Escanciano, Jacho-Chávez, and Lewbel (2016).

Finally, there are some general caveats regarding the application of any of these methods. Identification based on constructed instruments depends on strong modeling assumptions. So, when they're available, it is usually better to instead use 'true' outside instruments, that is, instruments that are known or believed to be excluded and exogenously determined based on randomization or on strong economic theory.

However, in practice one is often not sure if a candidate outside instrument is a valid instrument. A

candidate instrument might be invalid because the economic theory leading to its exclusion restriction is wrong. Even with randomization in a causal or experimental setting, assumptions like SUTVA (the stable unit treatment value assumption), or the assumption of no measurement error and no endogenous attrition could be violated, potentially making an instrument invalid despite randomization.

In these cases, constructed instruments can be used to provide overidentifying information for model tests and for robustness checks. In particular, one might estimate the model using both outside instruments and constructed instruments, and then test jointly for validity of all the instruments, using e.g., a Sargan (1958) and Hansen (1982) J-test. 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 in both the outside instrument and the constructed instrument. Both might then be used in estimation to maximize efficiency.

One could also just estimate the model separately using outside instruments and constructed instruments. If the estimates are similar across these different sets of identifying assumptions, then that provides support for the model and evidence that the results are not just artifacts of one particular set of identifying assumptions. More generally, identification based on functional form or 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 efficiency of estimation.

References

[1] Dong, Y. (2010), "Endogenous Regressor Binary Choice Models Without Instruments, With an Application to Migration," Economics Letters, 107(1), 33-35.

[2] Emran, M. S., V. Robano, and S. C. Smith, (2014) "Assessing the Frontiers of Ultrapoverty Reduction: Evidence from Challenging the Frontiers of Poverty Reduction/Targeting the Ultra-poor,

an Innovative Program in Bangladesh," Economic Development and Cultural Change, 62(2), pp. 339-380.

[3] Escanciano, J-C, D. Jacho-Chávez, and A. Lewbel (2016), “Identification and Estimation of Semiparametric Two Step Models,” Quantitative Economics, 7, pp. 561-589.

[4] Hansen, L. P. (1982). "Large Sample Properties of Generalized Method of Moments Estimators". Econometrica 50(4), pp. 1029–1054.

[5] Hoang, T. X., C. S. Pham, and M. A. Ulubaşoğlu, (2014) "Non-Farm Activity, Household Expenditure, and Poverty Reduction in Rural Vietnam: 2002–2008," World Development, 64, pp. 554–568.

[6] Klein, R. and F. Vella (2010), "Estimating a class of triangular simultaneous equations models without exclusion restrictions," Journal of Econometrics 154(2), pp. 154–164.

[7] Le Moglie, M., L. Mencarini, and C. Rapallini, (2015), "Is it just a matter of personality? On the role of subjective well-being in childbearing behavior," Journal of Economic Behavior and Organization, 117, pp. 453–475.

[8] Lewbel, A. (2012), “Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models.” Journal of Business and Economic Statistics, 30, 67-80.

[9] Lewbel, A. (2016), “The Identification Zoo - Meanings of Identification in Econometrics," Unpublished Mauscript, Boston College.

[10] Lewbel, A., Y. Dong, and T. T. Yang (2012), “Viewpoint: Comparing Features of Convenient Estimators for Binary Choice Models With Endogenous Regressors,” Canadian Journal of Economics, 45, 809-829.

[11] 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), pp. 800-824.

[12] Rigobon, R. (2003), ”Identification Through Heteroskedasticity,” Review of Economics and Statistics, 85(4), pp. 777-792.

[13] Sargan, J. D. (1958), "The Estimation of Economic Relationships Using Instrumental Variables". Econometrica 26(3), pp. 393–415.