Implementing new econometric tools in Stata

Christopher F Baum

Boston College and DIW Berlin

Mexican Stata Users Group Meeting, May 2013

Introduction

Focus of the talk: the implementation of two state-of-the-art econometric estimators in Stata and Mata.

1. estimating a binary response with one or more limited endogenous variables: `sspecialreg`.
2. estimating an equation with instrumental variables techniques where sufficient instruments may not be available: `ivreg2h`.

Topic I: Modeling a binary outcome with an endogenous binary regressor

Acknowledgement
This presentation is based on the work of Lewbel, Dong & Yang, “Viewpoint: Comparing features of Convenient Estimators for Binary Choice Models With Endogenous Regressors”, *Canadian Journal of Economics*, 45:3, 2012. Baum’s contribution is the review and enhancement of the software developed in this research project.

Motivation

- Researchers often want to estimate a binomial response, or binary choice, model where one or more explanatory variables are endogenous or mismeasured.
- For instance, in policy analysis, we want to estimate treatment effects when treatment is not randomly assigned, or based solely on the observables.
- A linear 2SLS model, equivalent to a linear probability model with instrumental variables, is often employed, ignoring the binary outcome.

Several alternative approaches exist:

- linear probability model (LPM) with instruments
- maximum likelihood estimation
- control function based estimation
- ‘special regressor’ methods

Each of these estimators has advantages and disadvantages, and some of these disadvantages are rarely acknowledged.

In what follows, we focus on a particular disadvantage of the linear probability model, and propose a straightforward alternative based on ‘special regressor’ methods (Lewbel, *J. Econometrics*, 2000; Lewbel, Dong and Yang, *Can. J. Econ.*, 2012).

We also propose the *average index function* (AIF), an alternative to the average structural function (ASF; Blundell and Powell, *Rev. Ec. Stud.*, 2004), for calculating marginal effects. It is easy to construct and estimate, as we will illustrate.

Binary choice models

We define D as an observed binary variable: the outcome to be explained. Let X be a vector of observed regressors, and β a corresponding coefficient vector, with ε an unobserved error. In a treatment model, X would include a binary treatment indicator T. In general, X could be divided into X^e, possibly correlated with ε, and X^0, which are exogenous.

A binary choice or ‘threshold crossing’ model estimated by maximum likelihood is

$$D = I(X\beta + \varepsilon \geq 0)$$

where $I(\cdot)$ is the indicator function. This latent variable approach is that employed in a binomial probit or logit model, with Normal or logistic errors, respectively. Although estimation provides point and interval estimates of β, the choice probabilities and marginal effects are of interest: that is, $\Pr[D = 1|X]$ and $\partial \Pr[D = 1|X]/\partial X$.

Linear probability models

In contrast to the threshold crossing latent variable approach, a linear probability model (LPM) assumes that

$$D = X\beta + \varepsilon$$

so that the estimated coefficients $\hat{\beta}$ are themselves the marginal effects. With all exogenous regressors, $E(D|X) = \Pr[D = 1|X] = X\beta$.

If some elements of X (possibly including treatment indicators) are endogenous or mismeasured, they will be correlated with ε. In that case, an instrumental variables approach is called for, and we can estimate the LPM with 2SLS or IV-GMM, given an appropriate set of instruments Z.

As the LPM with exogenous explanatory variables is based on standard regression, the zero conditional mean assumption $E(\varepsilon|X) = 0$ applies. In the presence of endogeneity or measurement error, the corresponding assumption $E(\varepsilon|Z) = 0$ applies, with Z the set of instruments, including the exogenous elements of X.

An obvious flaw in the LPM: the error ε cannot be independent of *any* regressors, even exogenous regressors, unless X consists of a single binary regressor. This arises because for any given X, ε must equal either $1 - X\beta$ or $-X\beta$, which are functions of all elements of X.

The other, well recognized, flaw in the LPM is that its fitted values are not constrained to lie in the unit interval, so that predicted probabilities below zero or above one are commonly encountered. Any regressor that can take on a large range of values will inevitably cause the LPM's predictions to breach these bounds.

A common rejoinder to these critiques is that the LPM is only intended to approximate the true probability for a limited range of X values, and that its constant marginal effects are preferable to those of the binary probit or logit model, which are functions of the values of all elements of X.

Consider, however, the LPM with a single continuous regressor. The linear prediction is an approximation to the *S*-shape of any cumulative distribution function: for instance, that of the Normal for the probit model. The linear prediction departs greatly from the *S*-shaped CDF long before it nears the (0,1) limits. Thus, the LPM will produce predicted probabilities that are too extreme (closer to zero or one) even for moderate values of $X\hat{\beta}$ that stay ‘in bounds’.

Some researchers claim that although predicted probabilities derived from the LPM are flawed, their main interest lies in the models’ marginal effects, and argue that it makes little substantive difference to use a LPM, with its constant marginal effects, rather than the more complex marginal effects derived from a proper estimated CDF, such as that of the probit model.

Example 1

Jeffrey Wooldridge's widely used undergraduate text, *Introductory Econometrics: A Modern Approach* devotes a section of the chapter on regression with qualitative variables to the LPM. He points out two flaws: computation of the predicted probability and marginal effects—and goes on to state

> *"Even with these problems, the linear probability model is useful and often applied in economics. It usually works well for values of the independent variables that are near the averages in the sample." (2009, p. 249)*

Wooldridge also discusses the heteroskedastic nature of the LPM's error, which is binomial by construction, but does not address the issue of the lack of independence that this implies.

EXAMPLE 2

Joshua Angrist and Steve Pischke's popular *Mostly Harmless Econometrics* invokes the principle of Occam's razor, arguing that

> *"...extra complexity comes into the inference step as well, since we need standard errors for marginal effects." (2009, p. 107)*

This is surely a red herring for Stata users, as the `margins` command in Stata 11 or 12 computes those standard errors via the delta method. They also discuss the difficulty of computing marginal effects for a binary regressor: again, not an issue for Stata 12 users, with the new `contrast` command.

Maximum likelihood estimators

A maximum likelihood estimator of a binary outcome with possibly endogenous regressors can be implemented for the model

$$\begin{aligned} D &= I(X^e \beta_e + X^0 \beta_0 + \varepsilon \geq 0) \\ X^e &= G(Z, \theta, e) \end{aligned}$$

which, for a single binary endogenous regressor, $G(\cdot)$ probit, and ε and e jointly Normal, is the model estimated by Stata's `biprobit` command.

Like the LPM, maximum likelihood allows endogenous regressors in X^e to be continuous, discrete, limited, etc. as long as a model for $G(\cdot)$ can be fully specified, along with the fully parameterized joint distribution of (ε, e).

Control function estimators

Control function estimators first estimate the model of endogenous regressors as a function of instruments, like the ‘first stage’ of 2SLS, then use the errors from this model as an additional regressor in the main model.

This approach is more general than maximum likelihood as the first stage function can be semiparametric or nonparametric, and the joint distribution of (ε, e) need not be fully parameterized.

To formalize the approach, consider a model $D = M(X, \beta, \varepsilon)$, and assume there are functions G, h and a well-behaved error U such that $X^e = G(Z, e), \varepsilon = h(e, U)$, and $U \perp (X, e)$.

We first estimate $G(\cdot)$: the endogenous regressors as functions of instruments Z, and derive fitted values of the errors e. Then we have

$$D = M(X, \beta, h(e, u)) = \widetilde{M}(X, e, \beta, U)$$

where the error term of the $\widetilde{M}$ model is U, which is suitably independent of (X, e). This model no longer has an endogeneity problem, and can be estimated via straightforward methods.

Given the threshold crossing model

$$\begin{aligned} D &= I(X^e\beta_e + X^0\beta_0 + \varepsilon \geq 0) \\ X^e &= Z\alpha + e \end{aligned}$$

with (ε, e) jointly normal, we can first linearly regress X^e on Z, with residuals being estimates of e.

This then yields an ordinary probit model

$$D = I(X^e\beta_e + X^0\beta_0 + \lambda e + U \geq 0)$$

which is the model estimated by Stata's `ivprobit` command. Despite its name, `ivprobit` is a control function estimator, not an IV estimator.

A substantial limitation of control function methods in this context is that they generally require the endogenous regressors X^e to be continuous, rather than binary, discrete, or censored. For instance, a binary endogenous regressor will violate the assumptions necessary to derive estimates of the ‘first stage’ error term e. The errors in the ‘first stage’ regression cannot be normally distributed and independent of the regressors. Thus, the `ivprobit` command should not be applied to binary endogenous regressors, as its documentation clearly states.

In this context, control function estimators—like maximum likelihood estimators—of binary outcome models require that the first stage model be correctly specified. This is an important limitation of these approaches. A 2SLS approach will lose efficiency if an appropriate instrument is not included, but a ML or control function estimator will generally become inconsistent.

Special regressor estimators

Special regressor estimators were first proposed by Lewbel (*J. Econometrics*, 2000). Their implementation is fully described in Dong and Lewbel (2012, BC WP 604; forthcoming, *Econometric Reviews*).

They assume that the model includes a particular regressor, V, with certain properties. It is exogenous (that is, $E(\varepsilon | V) = 0$) and appears as an additive term in the model. It is continuously distributed and has a large support.

A third condition, important for the success of this approach, is that V have a thick-tailed distribution. A regressor with greater kurtosis will be more useful as a special regressor, and one that is strictly normally distributed would lack the tail thickness to perform well.

The binary choice special regressor proposed by Lewbel (2000) has the ‘threshold crossing’ form

$$D = I(X^e\beta_e + X^0\beta_0 + V + \varepsilon \geq 0)$$

or, equivalently,

$$D = I(X\beta + V + \varepsilon \geq 0)$$

This is the same basic form for *D* as in the ML or control function (CF) approach. Note, however, that the special regressor *V* has been separated from the other exogenous regressors, and its coefficient normalized to unity: a harmless normalization.

Given a special regressor V, the only other requirements are those applicable to linear 2SLS: to handle endogeneity, the set of instruments Z must satisfy $E(\varepsilon|Z) = 0$, and $E(Z'X)$ must have full rank.

The main drawback of this method is that the special regressor V must be conditionally independent of ε. Even if it is exogenous, it could fail to satisfy this assumption because of the way in which V might affect other endogenous regressors. Also, V must be continuously distributed after conditioning on the other regressors, so that a term like V^2 could not be included as an additional regressor.

Apart from these restrictions on V, the special regressor (SR) method has none of the drawbacks of the three models discussed earlier:

- Unlike the LPM, the SR predictions stay 'in bounds' and is consistent with other threshold crossing models.
- Unlike ML and CF methods, the SR model does not require correct specification of the 'first stage' model: any valid set of instruments may be used, with only efficiency at stake.
- Unlike ML, the SR method has a linear form, not requiring iterative search.
- Unlike CF, the SR method can be used when endogenous regressors X^e are discrete or limited; unlike ML, there is a single estimation method, regardless of the characteristics of X^e.
- Unlike ML, the SR method permits unknown heteroskedasticity in the model errors.

The special regressor method imposes far fewer assumptions on the distribution of errors—particularly the errors e in the ‘first stage’ equations for X^e—than do CF or ML estimation methods. Therefore, SR estimators will be less efficient than these alternatives when the alternatives are consistent.

SR estimators may be expected to have larger standard errors and lower precision than other methods, *when those methods are valid*. However, if a special regressor V can be found, the SR method will be valid under much more general conditions than the ML and CF methods.

The average index function (AIF)

Consider the original estimation problem

$$D = I(X\beta + \varepsilon \geq 0)$$

where with generality one of the elements of X may be a special regressor V, with coefficient one. If ε is independent of X, the *propensity score* or *choice probability* is $\Pr[D = 1|X] = E(D|X) = E(D|X\beta) = F_{-\varepsilon}(X\beta) = \Pr(-\varepsilon \leq X\beta)$, with $F_{-\varepsilon}(\cdot)$ the probability distribution function of $-\varepsilon$. In the case of independent errors, these measures are identical.

When some regressors are endogenous, or generally when the assumption $X \perp \varepsilon$ is violated (e.g., by heteroskedasticity), these expressions may differ from one another.

Blundell and Powell (*Rev. Ec. Stud.*, 2004) propose using the average structural function (ASF) to summarize choice probabilities: $F_{-\varepsilon}(X\beta)$, even though ε is no longer independent of X. In this case, $F_{-\varepsilon|X}(X\beta|X)$ should be computed: a formidable task.

Lewbel, Dong and Yang (*Can. J. Econ.*, 2012) propose using the measure $E(D|X\beta)$, which they call the *average index function* (AIF), to summarize choice probabilities.

Like the ASF, the AIF is based on the estimated index, and equals the propensity score when $\varepsilon \perp X$. However, when this assumption is violated (by endogeneity or heteroskedasticity), the AIF is usually easier to estimate, via a unidimensional nonparametric regression of D on $X\beta$.

The AIF can be considered a middle ground between the propensity score and the ASF, as the former conditions on all covariates using $F_{-\varepsilon|X}$; the ASF conditions on no covariates using $F_{-\varepsilon}$; and the AIF conditions on the *index* of covariates, $F_{-\varepsilon|X\beta}$.

Define the function $M(X\beta) = E(D|X\beta)$, with derivatives m. The marginal effects of the regressors on the choice probabilities, as measured by the AIF, are $\partial E(D|X\beta)/\partial X = m(X\beta)\beta$, so the average marginal effects just equal the average derivatives, $E(m(X\beta + V))\beta$.

For the LPM, the ASF and AIF both equal the fitted values of the linear 2SLS regression of D on X. For the other methods, the AIF choice probabilities can be estimated using a standard unidimensional kernel regression of D on $X\hat{\beta}$: for instance, using the `lpoly` command in Stata, with the `at()` option specifying the observed data points. This will produce the AIF for each observation i, $\widehat{M}_i$.

Employing the derivatives of the kernel function, the individual-level marginal effects $\widehat{m}_i$ may be calculated, and averaged to produce average marginal effects:

$$\overline{m}\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\widehat{m}_i\hat{\beta}$$

Estimates of the precision of these average marginal effects may be derived by bootstrapping.

The Stata implementation

My Stata command `sspecialreg` estimates the Lewbel and Dong simple special regression estimator of a binary outcome with one or more binary endogenous variables. It is an optimized version of earlier code developed for this estimator, and provides significant (8–10x) speed improvements over that code.

Two forms of the special regressor estimator are defined, depending on assumptions made about the distribution of the special regressor V. In the first form of the model, only the mean of V is assumed to be related to the other covariates. In the second, 'heteroskedastic' form, higher moments of V can also depend in arbitrary, unknown ways on the other covariates. In practice, the latter form may include squares and cross products of some of the covariates in the estimation process, similar to the auxiliary regression used in White's general test for heteroskedasticity.

The `sspecialreg` Stata command also allows for two specifications of the density estimator used in the model: one based on a standard kernel density approach such as that implemented by `density` or Ben Jann's `kdens`, as well as the alternative 'sorted data density' approach proposed by Lewbel and Schennach (*J. Econometrics*, 2007). Implementation of the latter approach also benefited greatly, in terms of speed, by being rewritten in Mata, with Ben Jann's help gratefully acknowledged.

Just as in a `probit` or `ivprobit` model, the quantities of interest are not the estimated coefficients derived in the special regressor method, but rather the marginal effects. In the work of Lewbel et al., those are derived from the average index function (AIF) as described earlier. Point estimates of the AIF can be derived in a manner similar to that of average marginal effects in standard limited dependent variable models. For interval estimates, bootstrapped standard errors for the marginal effects are computed.

A bootstrap option was also added to `sspecialreg` so that the estimator can produce point and interval estimates of the relevant marginal effects in a single step, with the user's choice of the number of bootstrap samples to be drawn.

Empirical illustration 1

In this example of the special regressor method, taken from Dong and Lewbel (BC WP 604), the binary dependent variable is an indicator that individual *i* migrates from one US state to another. The objective is to estimate the probability of interstate migration.

The special regressor V_i in this context is age. Human capital theory suggests that it should appear linearly (or at least monotonically) in a threshold crossing model. Migration is in part driven by maximizing expected lifetime income, and the potential gain in lifetime earnings from a permanent change in labor income declines linearly with age. Evidence of empirical support for this relationship is provided by Dong (*Ec. Letters*, 2010). V_i is defined as the negative of age, demeaned, so that it should have a positive coefficient and a zero mean.

Pre-migration family income and home ownership are expected to be significant determinants of migration, and both should be considered endogenous. A maximum likelihood approach would require an elaborate dynamic specification in order to model the homeownership decision. Control function methods such as `ivprobit` are not appropriate as `homeowner` is a discrete variable.

The sample used includes male heads of household, 23–59 years of age, from the 1990 wave of the PSID who have completed education and are not retired, so as to exclude those moving to retirement communities. The observed $D = 1$ indicates migration during 1991–1993. In the sample of 4689 individuals, 807 were interstate migrants.

Exogenous regressors in the model include years of education, number of children, and indicators for white, disabled, and married individuals. The instruments Z also include the level of government benefits received in 1989–1990 and state median residential tax rates.

In the following table, we present four sets of estimates of the marginal effects computed by `ssimplereg`, utilizing the sorted data density estimator in columns 2 and 4 and allowing for heteroskedastic errors in columns 3 and 4.

For contrast, we present the results from an IV LPM (`ivregress 2sls`) in column 5, a standard `probit` (ignoring endogeneity) in column 6, and an `ivprobit` in the last column, ignoring its lack of applicability to the binary endogenous regressor `homeowner`.

Table: Marginal effects: binary outcome, binary endogenous regressor

	kdens	sortdens	kdens_hetero	sortdens_hetero	IV-LPM	probit	ivprobit
age	0.0146	0.0112	0.0071	0.0104	-0.0010	0.0019	-0.0005
	(0.003)***	(0.003)***	(0.003)*	(0.003)***	(0.002)	(0.001)**	(0.007)
log income	-0.0079	0.0024	0.0382	0.0176	0.0550	-0.0089	0.1406
	(0.028)	(0.027)	(0.024)	(0.026)	(0.080)	(0.007)	(0.286)
homeowner	0.0485	-0.0104	-0.0627	-0.0111	-0.3506	-0.0855	-1.0647
	(0.072)	(0.065)	(0.059)	(0.061)	(0.204)	(0.013)***	(0.708)
white	0.0095	0.0021	0.0021	0.0011	0.0086	-0.0099	0.0134
	(0.008)	(0.010)	(0.007)	(0.008)	(0.018)	(0.012)	(0.065)
disabled	0.1106	0.0730	0.0908	0.0916	0.0114	-0.0122	0.0104
	(0.036)**	(0.042)	(0.026)***	(0.037)*	(0.055)	(0.033)	(0.203)
education	-0.0043	-0.0023	-0.0038	-0.0036	0.0015	0.0004	0.0047
	(0.002)*	(0.003)	(0.002)*	(0.002)	(0.004)	(0.002)	(0.015)
married	0.0628	0.0437	0.0258	0.0303	0.0322	-0.0064	0.0749
	(0.020)**	(0.028)	(0.013)	(0.020)	(0.031)	(0.017)	(0.114)
nr. children	-0.0169	-0.0117	0.0006	-0.0021	0.0137	0.0097	0.0502
	(0.005)***	(0.005)*	(0.002)	(0.003)	(0.006)*	(0.005)*	(0.023)*

Note: bootstrapped standard errors in parentheses (100 replications)

The standard errors of these estimated marginal effects are computed from 100 bootstrap replications. The marginal effect of the ‘special regressor’ age of head is estimated as positive by the special regressor methods, but both the two-stage linear probability model and the ivprobit model yield negative (but insignificant) point estimates.

Household income and homeownership status do not seem to play significant roles in the migration decision. Among the special regression methods, the kernel data density estimator appears to yield the most significant results, with age of head, disabled status, years of education, marital status and number of children all playing a role in predicting the migration decision.

Empirical illustration 2

In this example of the special regressor method for firm-level data, we use a subsample of COMPUSTAT firm-level data on US publicly traded firms. The binary outcome variable is an indicator of whether the firm repurchased its common or preferred stock in a given year. The endogenous binary regressor is an indicator of whether the firm issued long-term debt in that year.

The special regressor V_{it} in this context is the firm's return on assets, or ROA. We also include the lagged value of income over total assets and a set of year dummies as exogenous factors. The instruments Z also include the lagged values of two ratios: capital expenditures to total assets and acquisitions to total assets.

The sample includes firm-level data from 1996–2006: a total of 30,852 firm-years. Over this period, the probability that a firm would repurchase its own stock in a given year is 0.479, while the probability that a firm would issue long-term debt in a given year is 0.583.

In the following table, we present four sets of estimates of the marginal effects computed by `sspecialreg`, utilizing the sorted data density estimator in columns 2 and 4 and allowing for heteroskedastic errors in columns 3 and 4.

For contrast, we present the results from an IV LPM (`ivregress 2sls`) in column 5, a standard `probit` (ignoring endogeneity) in column 6, and an `ivprobit` in the last column, ignoring its lack of applicability to the binary endogenous regressor `ltdiss`.

Modeling firms’ stock repurchase decision

Table: Marginal effects: binary outcome, binary endogenous regressor

	kdens	sortdens	kdens_H	sortdens_H	IV-LPM	probit	ivprobit
roa	0.4352	0.3918	0.3784*	0.3569	0.0122***	0.0065	0.0243**
	(0.231)	(0.307)	(0.189)	(0.284)	(0.002)	(0.004)	(0.008)
ltdiss	-0.1944***	-0.0847	-0.0983**	-0.0892	-0.4600***	0.0063	-1.5950***
	(0.025)	(0.092)	(0.034)	(0.082)	(0.058)	(0.007)	(0.067)
linca	-0.0246	-0.0098	-0.0345*	-0.0377	0.0794***	0.1644***	0.3687***
	(0.013)	(0.016)	(0.014)	(0.033)	(0.007)	(0.012)	(0.028)

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The standard errors of these estimated marginal effects are computed from 20 bootstrap replications. The marginal effect of the ‘special regressor’, ROA, is estimated as positive and significant by the special regressor methods.

The endogenous factor, long-term debt issuance, has a negative and significant coefficient (at 95%) in the estimates computed by the special regressor methods when the kernel density estimator is employed. The estimate of that coefficient in the IV-LPM model is almost twice as large, and its estimate from `ivprobit` implausibly huge.

The lagged ratio of income to total assets has a negative point estimate, varying in significance. There are sizable time effects (not shown) in the estimates.

Among the special regression methods, the kernel data density estimator allowing for heteroskedasticity appears to yield the most significant results, with all three variables displaying plausible estimates.

Summary remarks on `sspecialreg`

We have discussed an alternative to the linear probability model for estimation of a binary outcome with one or more binary endogenous regressors. This alternative, Lewbel and Dong's 'simple special regressor' method, circumvents the drawbacks of the IV-LPM approach, and yields consistent estimates in this context in which `ivprobit` does not.

Computation of marginal effects via the proposed average index function approach is straightforward, requiring only a single kernel density estimation and no iterative techniques. Bootstrapping is employed to derive interval estimates.

Topic II: IV methods with generated instruments

Acknowledgement
This presentation is based on the work of Arthur Lewbel, “Using Heteroskedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models,” *Journal of Business & Economic Statistics*, 2012. The contributions of Baum and Mark E Schaffer are the development of Stata software to implement Lewbel’s methodology.

Motivation

Instrumental variables (IV) methods are employed in linear regression models, e.g., $y = \mathbf{X}\beta + u$, where violations of the zero conditional mean assumption $\mathrm{E}[u|\mathbf{X}] = 0$ are encountered.

Reliance on IV methods usually requires that appropriate instruments are available to identify the model: often via exclusion restrictions.

Those instruments, $\mathbf{Z}$, must satisfy three conditions: (i) they must themselves satisfy orthogonality conditions ($\mathrm{E}[u\mathbf{Z}] = 0$); (ii) they must exhibit meaningful correlations with $\mathbf{X}$; and (iii) they must be properly excluded from the model, so that their effect on the response variable is only indirect.

Finding appropriate instruments which simultaneously satisfy all three of these conditions is often problematic, and the major obstacle to the use of IV techniques in many applied research projects.

Although textbook treatments of IV methods stress their usefulness in dealing with endogenous regressors, they are also employed to deal with omitted variables, or with measurement error of the regressors (‘errors in variables’) which if ignored will cause bias and inconsistency in OLS estimates.

Lewbel's approach

The method proposed in Lewbel (*JBES*, 2012) serves to identify structural parameters in regression models with endogenous or mismeasured regressors in the absence of traditional identifying information, such as external instruments or repeated measurements.

Identification is achieved in this context by having regressors that are uncorrelated with the product of heteroskedastic errors, which is a feature of many models where error correlations are due to an unobserved common factor.

In this presentation, we describe a method for constructing instruments as simple functions of the model's data. This approach may be applied when no external instruments are available, or, alternatively, used to supplement external instruments to improve the efficiency of the IV estimator.

Supplementing external instruments can also allow 'Sargan–Hansen' tests of the orthogonality conditions to be performed which would not be available in the case of exact identification by external instruments.

In that context, the approach is similar to the dynamic panel data estimators of Arellano and Bond (*Review of Economic Studies*, 1991) et al., as those estimators customarily make use of appropriate lagged values of endogenous regressors to identify the model. In contrast, the approach we describe here may be applied in a purely cross-sectional context, as well as that of time series or panel data.

The basic framework

Consider Y_1, Y_2 as observed endogenous variables, X a vector of observed exogenous regressors, and $\varepsilon = (\varepsilon_1, \varepsilon_2)$ as unobserved error processes. Consider a structural model of the form:

$$\begin{aligned} Y_1 &= X'\beta_1 + Y_2\gamma_1 + \varepsilon_1 \\ Y_2 &= X'\beta_2 + Y_1\gamma_2 + \varepsilon_2 \end{aligned}$$

This system is triangular when $\gamma_2 = 0$ (or, with renumbering, when $\gamma_1 = 0$). Otherwise, it is fully simultaneous. The errors $\varepsilon_1, \varepsilon_2$ may be correlated with each other.

If the exogeneity assumption, $E(\varepsilon X) = 0$ holds, the reduced form is identified, but in the absence of identifying restrictions, the structural parameters are not identified. These restrictions often involve setting certain elements of β_1 or β_2 to zero, which makes instruments available.

In many applied contexts, the third assumption made for the validity of an instrument—that it only indirectly affects the response variable—is difficult to establish. The zero restriction on its coefficient may not be plausible. The assumption is readily testable, but if it does not hold, IV estimates will be inconsistent.

Identification in Lewbel's approach is achieved by restricting correlations of $\varepsilon\varepsilon'$ with X. This relies upon higher moments, and is likely to be less reliable than identification based on coefficient zero restrictions. However, in the absence of plausible identifying restrictions, this approach may be the only reasonable strategy.

The parameters of the structural model will remain unidentified under the standard homoskedasticity assumption: that $\mathrm{E}(\varepsilon\varepsilon' | X)$ is a matrix of constants. However, in the presence of heteroskedasticity related to at least some elements of X, identification can be achieved.

In a fully simultaneous system, assuming that $\mathrm{cov}(X, \varepsilon_j^2) \neq 0, j = 1, 2$ and $\mathrm{cov}(Z, \varepsilon_1 \varepsilon_2) = 0$ for observed Z will identify the structural parameters. Note that Z may be a subset of X, so no information outside the model specified above is required.

The key assumption that $\mathrm{cov}(Z, \varepsilon_1 \varepsilon_2) = 0$ will automatically be satisfied if the mean zero error processes are conditionally independent: $\varepsilon_1 \perp \varepsilon_2 | Z = 0$. However, this independence is not strictly necessary.

Single-equation estimation

In the most straightforward context, we want to apply the instrumental variables approach to a single equation, but lack appropriate instruments or identifying restrictions. The auxiliary equation or ‘first-stage’ regression may be used to provide the necessary components for Lewbel’s method.

In the simplest version of this approach, generated instruments can be constructed from the auxiliary equations’ residuals, multiplied by each of the included exogenous variables in mean-centered form:

$$Z_j = (X_j - \overline{X}) \cdot \epsilon$$

where ϵ is the vector of residuals from the ‘first-stage regression’ of each endogenous regressor on all exogenous regressors, including a constant vector.

These auxiliary regression residuals have zero covariance with each of the regressors used to construct them, implying that the means of the generated instruments will be zero by construction. However, their element-wise products with the centered regressors will not be zero, and will contain sizable elements if there is clear evidence of ‘scale heteroskedasticity’ with respect to the regressors. Scale-related heteroskedasticity may be analyzed with a Breusch–Pagan type test: `estat hettest` in an OLS context, or `ivhettest` (Schaffer, SSC; Baum et al., *Stata Journal*, 2007) in an IV context.

The greater the degree of scale heteroskedasticity in the error process, the higher will be the correlation of the generated instruments with the included endogenous variables which are the regressands in the auxiliary regressions.

Stata implementation

An implementation of this simplest version of Lewbel's method, `ivreg2h`, has been constructed from Baum, Schaffer, Stillman's `ivreg2` and Schaffer's `xtivreg2`, both available from the SSC Archive. The panel-data features of `xtivreg2` are not used in this implementation: only the nature of `xtivreg2` as a 'wrapper' for `ivreg2`.

In its current version, `ivreg2h` can be invoked to estimate

- a traditionally identified single equation, or
- a single equation that fails the order condition for identification: either (i) by having no excluded instruments, or
(ii) by having fewer excluded instruments than needed for traditional identification.

In the former case, of external instruments augmented by generated instruments, the program provides three sets of estimates: the traditional IV estimates, estimates using only generated instruments, and estimates using both generated and excluded instruments.

In the latter case, of an underidentified equation, only the estimates using generated instruments are displayed. Unlike `ivreg2` or `ivregress`, `ivreg2h` allows the syntax

```
ivreg2h depvar exogvar (endogvar=)
```

disregarding the failure of the order condition for identification.

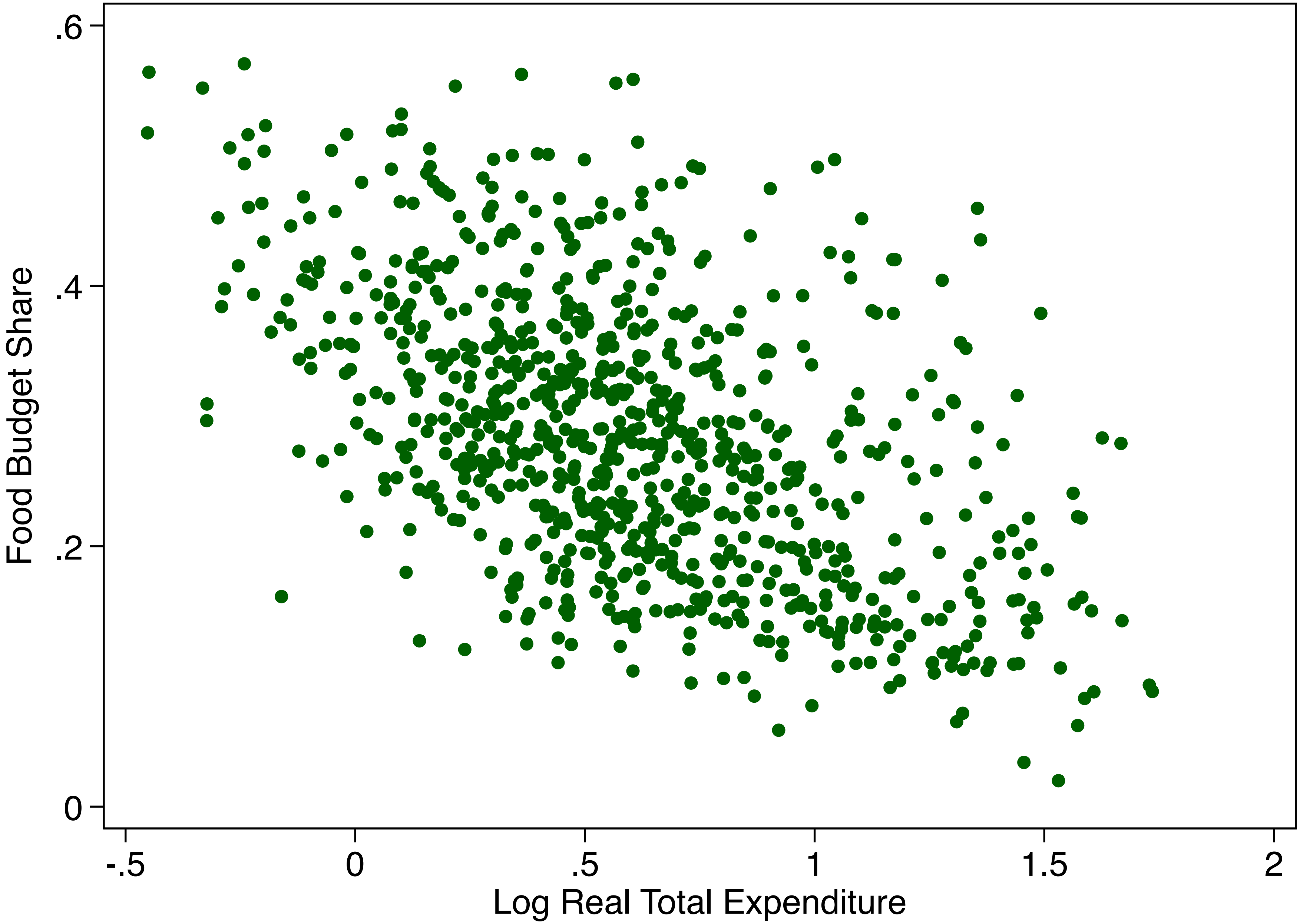
Empirical illustration 1

In Lewbel's 2012 *JBES* paper, he illustrates the use of his method with an Engel curve for food expenditures. An Engel curve describes how household expenditure on a particular good or service varies with household income (Ernst Engel, 1857, 1895).[1] Engel's research gave rise to *Engel's Law*: while food expenditures are an increasing function of income and family size, food budget shares decrease with income (Lewbel, *New Palgrave Dictionary of Economics*, 2d ed. 2007).

In this application, we are considering a key explanatory variable, total expenditures, to be subject to potentially large measurement errors, as is often found in applied research: due in part to infrequently purchased items (Meghir and Robin, *Journal of Econometrics*, 1992).

[1] Not to be confused with Friedrich Engels, Karl Marx's coauthor.

The data are 854 households, all married couples without children, from the UK Family Expenditure Survey, 1980–1982, as studied by Banks, Blundell and Lewbel (*Review of Economics and Statistics*, 1997). The dependent variable is the food budget share, with a sample mean of 0.285. The key explanatory variable is log real total expenditures, with a sample mean of 0.599. A number of additional regressors (age, spouse's age, ages^2, and a number of indicators) are available as controls. The coefficients of interest in this model are those of log real total expenditures and the constant term.


Food Budget Share
.6
.4
.2
0
-.5
0
.5
1
1.5
2
Log Real Total Expenditure

We first estimate the model with OLS regression, ignoring any issue of mismeasurement. We then reestimate the model with log total income as an instrument using two-stage least squares: an exactly identified model. As such, this is also the IV-GMM estimate of the model.

In the following table, these estimates are labeled as OLS and TSLS1. A Durbin–Wu–Hausman test for the endogeneity of log real total expenditures in the TSLS1 model rejects with p-value=0.0203, indicating that application of OLS is inappropriate.

Table: OLS and conventional TSLS

	(1) OLS	(2) TSLS,ExactID
lrtotexp	-0.127	-0.0859
	(0.00838)	(0.0198)
Constant	0.361	0.336
	(0.00564)	(0.0122)

Standard errors in parentheses

These OLS and TSLS results can be estimated with standard `regress` and `ivregress 2sls` commands. We now turn to estimates produced from generated instruments via Lewbel's method.

We produce generated instruments from each of the exogenous regressors in this equation. The equation may be estimated by TSLS or by IV-GMM, in each case producing robust standard errors. For IV-GMM, we report Hansen's *J*.

Table: Generated instruments only

	(1) TSLS,GenInst	(2) GMM,GenInst
lrtotexp	-0.0554	-0.0521
	(0.0589)	(0.0546)
Constant	0.318	0.317
	(0.0352)	(0.0328)
Jval		12.91
Jdf		11
Jpval		0.299

Standard errors in parentheses

The greater efficiency available with IV-GMM is evident in the precision of these estimates. However, reliance on generated instruments yields much larger standard errors than identified TSLS.[2]

As an alternative, we augment the available instrument, log total income, with the generated instruments, which overidentifies the equation, estimated with both TSLS and IV-GMM methods.

[2] The GMM results do not agree with those labeled GMM2 in the *JBES* article. However, it appears that the published GMM2 results are not the true optimum.

Table: Augmented by generated instruments

	(1) TSLS,AugInst	(2) GMM,AugInst
lrtotexp	-0.0862	-0.0867
	(0.0186)	(0.0182)
Constant	0.336	0.337
	(0.0114)	(0.0112)
Jval		16.44
Jdf		12
Jpval		0.172

Standard errors in parentheses

Relative to the original, exactly-identified TSLS/IV-GMM specification, the use of generated instruments to augment the model has provided an increase in efficiency, and allowed overidentifying restrictions to be tested. As a comparison:

Table: With and without generated instruments

	(1) GMM,ExactID	(2) GMM,AugInst
lrtotexp	-0.0859	-0.0867
	(0.0198)	(0.0182)
Constant	0.336	0.337
	(0.0122)	(0.0112)
Jval		16.44
Jdf		12
Jpval		0.172

Standard errors in parentheses

Empirical illustration 2

We illustrate the use of this method with an estimated equation on firm-level panel data from US Industrial Annual COMPUSTAT. The model, a variant on that presented in a working paper by Baum, Chakraborty and Liu, is based on Faulkender and Wang (*J. Finance*, 2006).

We seek to explain the firm-level unexpected change in cash holdings as a function of the level of cash holdings (C), the change in earnings (dE), the change in non-cash assets (dNA) and market leverage (Lev). The full sample contains about 13,000 firms for up to 35 years.

For purposes of illustration, we first fit the model treating the level of cash holdings as endogenous, but maintaining that we have no available external instruments. In this context, `ivreg2h` produces three generated instruments: one from each included exogenous regressor. We employ IV-GMM with a cluster-robust VCE, clustered by firm.

Table: Modeling $\Delta C1$

	GenInst
C	-0.152***
	(-5.04)
dE	0.0301***
	(7.24)
dNA	-0.0115***
	(-6.00)
Lev	-0.0447***
	(-18.45)
N	117036
jdf	2
jp	0.245

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The resulting model is overidentified by two degrees of freedom (jdf). The jp value of 0.245 is the p-value of the Hansen *J* statistic.

We reestimate the model using the lagged value of cash holdings as an instrument. This causes the model to be exactly identified, and estimable with standard techniques. ivreg2h thus produces three sets of estimates: those for standard IV, those using only generated instruments, and those using both external and generated instruments. The generated-instrument results differ from those shown previously as the sample is now smaller.

Table: Modeling $\Delta C1$

	StdIV	GenInst	GenExtInst
C	-0.0999***	-0.127***	-0.100***
	(-15.25)	(-3.83)	(-15.37)
dE	0.0287***	0.0324***	0.0304***
	(6.43)	(7.09)	(7.99)
dNA	-0.0121***	-0.0133***	-0.0124***
	(-6.16)	(-6.43)	(-7.09)
Lev	-0.0447***	-0.0468***	-0.0460***
	(-15.99)	(-17.79)	(-19.51)
N	102870	102870	102870
jdf	0	2	3
jp		0.691	0.697

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The results show that there are minor differences in the point estimates produced by standard IV and those from the augmented equation. However, the latter are more efficient, with smaller standard errors for each coefficient. The model is now overidentified by three degrees of freedom, allowing us to conduct a test of over identifying restrictions. The p-value of that test, `jp`, indicates no problem.

This example illustrates what may be the most useful aspect of Lewbel's method: the ability to augment an exactly-identified equation to both allow a test of over identifying restrictions and gain efficiency.

We have illustrated this method with one endogenous regressor, but it generalizes to multiple endogenous (or mismeasured) regressors. It may be employed as long as there is at least one included exogenous regressor for each endogenous regressor. If there is only one, the resulting equation will be exactly identified.

As this estimator has been implemented within the `ivreg2` framework, all of the diagnostics and options available in that program (Baum, Schaffer, Stillman, *Stata Journal*, 2003, 2007) are available in this context.

Summary remarks on `ivreg2h`

The extension of this method to the panel fixed-effects context is relatively straightforward, and we are finalizing a version of Mark Schaffer's `xtivreg2` which implements Lewbel's method in this context.

We have illustrated how this method might be used to augment the available instruments to facilitate the use of tests of overidentification. Lewbel argues that the method might also be employed in a fully saturated model, such as a difference-in-difference specification with all feasible fixed effects included, in order to test whether OLS methods will yield consistent results.

Concluding remarks

I hope that this illustration of how cutting-edge econometric techniques may be made available to Stata users has been enlightening. The underlying code for both `sspecialreg` and `ivreg2h` is accessible, as is true of nearly all user-written contributions to the Stata community. This openness greatly enhances users' ability to both extend Stata's capabilities and assure themselves that user-written routines are of high quality and generally reliable.