The Stata Journal (2009)
9, Number 1, pp. 86–136

How to do xtabond2: An introduction to difference and system GMM in Stata

David Roodman
Center for Global Development
Washington, DC
droodman@cgdev.org

Abstract. The difference and system generalized method-of-moments estimators, developed by Holtz-Eakin, Newey, and Rosen (1988, *Econometrica* 56: 1371–1395); Arellano and Bond (1991, *Review of Economic Studies* 58: 277–297); Arellano and Bover (1995, *Journal of Econometrics* 68: 29–51); and Blundell and Bond (1998, *Journal of Econometrics* 87: 115–143), are increasingly popular. Both are general estimators designed for situations with "small T, large N" panels, meaning few time periods and many individuals; independent variables that are not strictly exogenous, meaning they are correlated with past and possibly current realizations of the error; fixed effects; and heteroskedasticity and autocorrelation within individuals. This pedagogic article first introduces linear generalized method of moments. Then it describes how limited time span and potential for fixed effects and endogenous regressors drive the design of the estimators of interest, offering Stata-based examples along the way. Next it describes how to apply these estimators with `xtabond2`. It also explains how to perform the Arellano–Bond test for autocorrelation in a panel after other Stata commands, using `abar`. The article concludes with some tips for proper use.

Keywords: st0159, xtabond2, generalized method of moments, GMM, Arellano–Bond test, abar

1 Introduction

Arellano–Bond (Arellano and Bond 1991) and Arellano–Bover/Blundell–Bond (Arellano and Bover 1995; Blundell and Bond 1998) dynamic panel estimators are increasingly popular. Both are general estimators designed for situations with 1) "small T, large N" panels, meaning few time periods and many individuals; 2) a linear functional relationship; 3) one left-hand-side variable that is dynamic, depending on its own past realizations; 4) independent variables that are not strictly exogenous, meaning they are correlated with past and possibly current realizations of the error; 5) fixed individual effects; and 6) heteroskedasticity and autocorrelation within individuals but not across them. Arellano–Bond estimation starts by transforming all regressors, usually by differencing, and uses the generalized method of moments (GMM) (Hansen 1982), and is called difference GMM. The Arellano–Bover/Blundell–Bond estimator augments Arellano–Bond by making an additional assumption that first differences of instrument variables are uncorrelated with the fixed effects. This allows the introduction of more instruments and can dramatically improve efficiency. It builds a system of two

equations—the original equation and the transformed one—and is known as system GMM.

The `xtabond2` command implements these estimators. When introduced in late 2003, it brought several novel capabilities to Stata users. Going beyond the built-in `xtabond` command, `xtabond2` implemented system GMM. It made the Windmeijer (2005) finite-sample correction to the reported standard errors in two-step estimation, without which those standard errors tend to be severely downward biased. It introduced finer control over the instrument matrix. And in later versions, it offered automatic difference-in-Sargan/Hansen testing for the validity of instrument subsets; support for observation weights; and the forward orthogonal deviations transform, an alternative to differencing proposed by Arellano and Bover (1995) that preserves sample size in panels with gaps. Stata 10 absorbed many of these features. `xtabond` now performs the Windmeijer correction. The new `xtdpd` and `xtdpdsys` commands jointly offer most of `xtabond2`'s features, while moving somewhat toward its syntax and running significantly faster. On the other hand, `xtabond2` runs in older versions of Stata and still offers unique features including observation weights, automatic difference-in-Sargan/Hansen testing, and the ability to "collapse" instruments to limit instrument proliferation.

Interestingly, though the Arellano and Bond article (1991) is now seen as the source of an estimator, it is entitled *Some tests of specification for panel data*. The instrument sets and use of GMM that largely define difference GMM originated with Holtz-Eakin, Newey, and Rosen (1988). One of Arellano and Bond's contributions is a test for autocorrelation appropriate for linear GMM regressions on panels, which is especially important when lags are used as instruments. `xtabond2` automatically reports this test. But because ordinary least squares (OLS) and two-stage least squares (2SLS) are special cases of linear GMM, the Arellano–Bond test has wider applicability. The postestimation command `abar`, also described in this article, makes the test available after `regress`, `ivregress`, `ivreg2`, `newey`, and `newey2`.

One disadvantage of difference and system GMM is that they are complicated and so can easily generate invalid estimates. Implementing them with a Stata command stuffs them into a black box, creating the risk that users not understanding the estimators' purpose, design, and limitations will unwittingly misuse the estimators. This article aims to prevent that misuse. Its approach is therefore pedagogic. Section 2 introduces linear GMM. Section 3 describes how certain panel econometric problems drive the design of the difference and system estimators. Some of the derivations are incomplete because their purpose is to build intuition; the reader must refer to the original article or to textbooks for details. Section 4 describes the `xtabond2` and `abar` syntaxes, with examples. Section 5 concludes the article with tips for proper use.

(*Continued on next page*)

2 Linear GMM[1]

2.1 The GMM estimator

The classical linear estimators OLS and 2SLS can be thought of in several ways, the most intuitive being suggested by the estimators' names. OLS minimizes the sum of the squared errors. 2SLS can be implemented via OLS regressions in two stages. But there is another, more unified, way to view these estimators. In OLS, identification can be said to flow from the assumption that the regressors are orthogonal to the errors; the inner products, or *moments*, of the regressors with the errors are set to 0. Likewise, in the more general 2SLS framework, which distinguishes between regressors and instruments while allowing the two categories to overlap (variables in both categories are *included, exogenous regressors*), the estimation problem is to choose coefficients on the regressors so that the moments of the errors with the instruments are 0.

However, an ambiguity arises in conceiving of 2SLS as a matter of satisfying such *moment conditions*. What if there are more instruments than regressors? If we view the moment conditions as a system of equations, one for each instrument, then the unknowns in these equations are the coefficients, of which there is one for each regressor. If instruments outnumber regressors, then equations outnumber unknowns and the system usually cannot be solved. Thus the moment conditions cannot be expected to hold perfectly in finite samples even when they are true asymptotically. This is the sort of problem we are interested in. To be precise, we want to fit the model

$$y = \mathbf{x}'\beta + \varepsilon$$
$$E(\varepsilon \,|\, \mathbf{z}) = 0$$

where β is a column vector of coefficients, y and ε are random variables, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)'$ is a column vector of k regressors, $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_j)'$ is column vector of j instruments, $\mathbf{x}$ and $\mathbf{z}$ can share elements, and $j \geq k$. We use $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ to represent matrices of N observations for $\mathbf{x}$, y, and $\mathbf{z}$, and we define $\mathbf{E} = \mathbf{Y} - \mathbf{X}\beta$. Given an estimate, $\widehat{\beta}$, the empirical residuals are $\widehat{\mathbf{E}} = (\widehat{e}_1, \ldots, \widehat{e}_N)' = \mathbf{Y} - \mathbf{X}\widehat{\beta}$. We make no assumption at this point about $E(\mathbf{E}\mathbf{E}' \,|\, \mathbf{Z}) = \mathbf{\Omega}$ except that it exists.

The challenge in estimating this model is that while all the instruments are theoretically orthogonal to the error term, $E(\mathbf{z}\varepsilon) = \mathbf{0}$, trying to force the corresponding vector of empirical moments, $E_N(\mathbf{z}\varepsilon) \equiv (1/N)\mathbf{Z}'\widehat{\mathbf{E}}$, to zero creates a system with more equations than variables if $j > k$. The specification is then *overidentified*. Because we cannot expect to satisfy all the moment conditions at once, the problem is to satisfy them all as well as possible in some sense, that is, to minimize the magnitude of the vector $E_N(\mathbf{z}\varepsilon)$.

1. For another introduction to GMM, see Baum, Schaffer, and Stillman (2003). For fuller accounts, see Ruud (2000, chap. 21–22) and Hayashi (2000, chap. 3).

In the GMM, one defines that magnitude through a generalized metric, based on a positive-semidefinite quadratic form. Let $\mathbf{A}$ be the matrix for such a quadratic form. Then the metric is

$$\| E_N(\mathbf{z}\varepsilon) \|_{\mathbf{A}} = \| \frac{1}{N}\mathbf{Z}'\widehat{\mathbf{E}} \|_{\mathbf{A}} \equiv N\left(\frac{1}{N}\mathbf{Z}'\widehat{\mathbf{E}}\right)' \mathbf{A}\left(\frac{1}{N}\mathbf{Z}'\widehat{\mathbf{E}}\right) = \frac{1}{N}\widehat{\mathbf{E}}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\widehat{\mathbf{E}} \quad (1)$$

To derive the implied GMM estimate, call it $\widehat{\beta}_{\mathbf{A}}$, we solve the minimization problem $\widehat{\beta}_{\mathbf{A}} = \text{argmin}_{\widehat{\beta}} \| \mathbf{Z}'\widehat{\mathbf{E}} \|_{\mathbf{A}}$, whose solution is determined by $\mathbf{0} = d/(d\widehat{\beta}) \| \mathbf{Z}'\widehat{\mathbf{E}} \|_{\mathbf{A}}$. Expanding this derivative with the chain rule gives

$$\begin{aligned}\mathbf{0} &= \frac{d}{d\widehat{\beta}} \| \mathbf{Z}'\widehat{\mathbf{E}} \|_{\mathbf{A}} = \frac{d}{d\widehat{\mathbf{E}}} \| \mathbf{Z}'\widehat{\mathbf{E}} \|_{\mathbf{A}} \frac{d\widehat{\mathbf{E}}}{d\widehat{\beta}} = \frac{d}{d\widehat{\mathbf{E}}} \left\{ \frac{1}{N}\widehat{\mathbf{E}}' \left(\mathbf{Z}\mathbf{A}\mathbf{Z}'\right) \widehat{\mathbf{E}} \right\} \frac{d\left(\mathbf{Y} - \mathbf{X}\widehat{\beta}\right)}{d\widehat{\beta}} \\ &= \frac{2}{N}\widehat{\mathbf{E}}'\mathbf{Z}\mathbf{A}\mathbf{Z}' \left(-\mathbf{X}\right)\end{aligned}$$

The last step uses the matrix identities $d\mathbf{A}\mathbf{b}/d\mathbf{b} = \mathbf{A}$ and $d\left(\mathbf{b}'\mathbf{A}\mathbf{b}\right)/d\mathbf{b} = 2\mathbf{b}'\mathbf{A}$, where $\mathbf{b}$ is a column vector and $\mathbf{A}$ is a symmetric matrix. Dropping the factor of $-2/N$ and transposing,

$$\begin{aligned}&\mathbf{0} = \widehat{\mathbf{E}}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X} = \left(\mathbf{Y} - \mathbf{X}\widehat{\beta}_{\mathbf{A}}\right)' \mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X} = \mathbf{Y}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X} - \widehat{\beta}'_{\mathbf{A}}\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X} \\ &\Rightarrow \mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\widehat{\beta}_{\mathbf{A}} = \mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{Y} \\ &\Rightarrow \widehat{\beta}_{\mathbf{A}} = \left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{Y} \quad (2)\end{aligned}$$

This is the GMM estimator defined by $\mathbf{A}$. It is linear in $\mathbf{Y}$.

While $\mathbf{A}$ weights *moments*, one can also incorporate weights for *observations*. If $\mathbf{W}$ is a diagonal $N \times N$ observation weighting matrix, then the GMM criterion function can be recast as $\| (1/N)\mathbf{Z}'\mathbf{W}\widehat{\mathbf{E}} \|_{\mathbf{A}}$. The appendix derives the more general weighted GMM estimator.

The GMM estimator is consistent, meaning that under appropriate conditions, it converges in probability to β as sample size goes to infinity (Hansen 1982). But like 2SLS, it is, in general, biased, as section 2.6 discusses, because in finite samples the instruments are almost always at least slightly correlated with the endogenous components of the instrumented regressors. Correlation coefficients between finite samples of uncorrelated variables are usually not exactly 0.

For future reference, the bias of the estimator is the corresponding projection of the true model errors:

$$\begin{aligned}\widehat{\beta}_{\mathbf{A}} - \beta &= \left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}' \left(\mathbf{X}\beta + \mathbf{E}\right) - \beta \\ &= \left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\beta + \left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{E} - \beta \\ &= \left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{E} \quad (3)\end{aligned}$$

2.2 Efficiency

It can be seen from (2) that multiplying **A** by a nonzero scalar would not change $\widehat{\beta}_{\mathbf{A}}$. But up to a factor of proportionality, each choice of **A** implies a different linear, consistent estimator of β. Which **A** should the researcher choose? Making **A** scalar is intuitive, generally inefficient, and instructive. By (1), it would yield an equal-weighted Euclidian metric on the moment vector. To see the inefficiency, consider what happens if there are two instruments of zero means, one drawn from a variable with a variance of 1, the other from a variable with a variance of 1,000. Moments based on the second instrument would easily dominate the minimization problem under equal weighting, wasting the information in the first. Or imagine a cross-country growth regression instrumenting with two highly correlated proxies for the poverty level. The marginal information content in the second would be minimal, yet including it in the moment vector would double the weight of poverty at the expense of other instruments. In both examples, the inefficiency is signaled by high variance or covariance among moments. This suggests that making **A** scalar is inefficient unless the moments $\mathbf{z}\varepsilon$ have equal variance and are uncorrelated—that is, if $\mathrm{Var}(\mathbf{z}\varepsilon)$ is itself scalar. This suggestion is correct, as will be seen.[2]

But that negative conclusion hints at the general solution. For efficiency, **A** must in effect weight moments in inverse proportion to their variances and covariances. In the first example above, such reweighting would appropriately deemphasize the high-variance instrument. In the second example, it would efficiently down-weight one or both of the poverty proxies. In general, for efficiency, we weight by the inverse of the variance of the population moments, which, under suitable conditions, is the asymptotic variance of the sample moments.

Because efficiency is an asymptotic notion, to speak rigorously of it we view matrices such as **Z** and **E** as elements of infinite sequences indexed by N. For economy of space, however, we suppress subscripts that would indicate this dependence. So we write the efficient GMM moment weighting matrix as

$$\mathbf{A}_{\mathrm{EGMM}} \equiv \mathrm{Var}(\mathbf{z}\varepsilon)^{-1} = \mathrm{Avar}\left(\frac{1}{N}\mathbf{Z}'\mathbf{E}\right)^{-1} \equiv \left\{\operatorname*{plim}_{N\to\infty} N\mathrm{Var}\left(\frac{1}{N}\mathbf{Z}'\mathbf{E}\right)\right\}^{-1} \tag{4}$$

Substituting into (1), the efficient GMM (EGMM) estimator minimizes

$$\| \mathbf{Z}'\widehat{\mathbf{E}} \|_{\mathbf{A}_{\mathrm{EGMM}}} = N\left(\frac{1}{N}\mathbf{Z}'\widehat{\mathbf{E}}\right)' \mathrm{Var}(\mathbf{z}\varepsilon)^{-1} \frac{1}{N}\mathbf{Z}'\widehat{\mathbf{E}}$$

Substituting this choice of **A** into (2) gives the direct formula for EGMM:

$$\widehat{\beta}_{\mathrm{EGMM}} = \left\{\mathbf{X}'\mathbf{Z}\mathrm{Var}(\mathbf{z}\varepsilon)^{-1}\mathbf{Z}'\mathbf{X}\right\}^{-1} \mathbf{X}'\mathbf{Z}\mathrm{Var}(\mathbf{z}\varepsilon)^{-1}\mathbf{Z}'\mathbf{Y} \tag{5}$$

EGMM is not feasible, however, unless $\mathrm{Var}(\mathbf{z}\varepsilon)$ is known.

2. This argument is analogous to that for the design of generalized least squares (GLS); GLS is derived with reference to the errors **E**, where GMM is derived with reference to the moments $\mathbf{Z}'\mathbf{E}$.

Before we move to making the estimator feasible, we demonstrate its efficiency. Define $\mathbf{S_{ZY}} = E_N(\mathbf{z}y) = (1/N)\mathbf{Z}'\mathbf{Y}$ and $\mathbf{S_{ZX}} = E_N(\mathbf{z}\mathbf{x}') = (1/N)\mathbf{Z}'\mathbf{X}$. We can then rewrite the general GMM estimator in (2) as $\widehat{\beta}_{\mathbf{A}} = (\mathbf{S}'_{\mathbf{ZX}}\mathbf{A}\mathbf{S}_{\mathbf{ZX}})^{-1}\mathbf{S}'_{\mathbf{ZX}}\mathbf{A}\mathbf{S}_{\mathbf{ZY}}$. We assume that conditions suitable for a Law of Large Numbers hold, so that

$$\mathbf{\Sigma_{ZX}} \equiv E(\mathbf{z}\mathbf{x}') = \operatorname*{plim}_{N\to\infty} \mathbf{S_{ZX}} \tag{6}$$

$$\begin{aligned}\operatorname*{plim}_{N\to\infty} N\mathrm{Var}(\mathbf{S_{ZY}}) &\equiv \mathrm{Avar}(\mathbf{S_{ZY}}) = \mathrm{Avar}\left(\frac{1}{N}\mathbf{Z}'\mathbf{Y}\right) = \mathrm{Avar}\left(\frac{1}{N}\mathbf{Z}'\mathbf{E}\right) \\ &= \mathrm{Var}(\mathbf{z}\varepsilon) = \mathbf{A}^{-1}_{\mathrm{EGMM}}\end{aligned} \tag{7}$$

For each sample size $N > 0$, let B_N be the vector space of scalar-valued functions of the random vector $\mathbf{Y}$. This space contains all the coefficient estimates defined by linear estimators based on $\mathbf{Y}$. For example, if $\mathbf{c} = (1\ 0\ 0\ldots)$, then $\mathbf{c}\widehat{\beta}_{\mathbf{A}} \in B_N$ is the estimated coefficient for x_1 according to the GMM estimator implied by some $\mathbf{A}$. We define an inner product on B_N by $\langle b_1, b_2\rangle = \mathrm{Cov}(b_1, b_2)$; the corresponding metric is $\| b \|^2 = \mathrm{Var}(b)$. The assertion that (5) is efficient is equivalent to saying that for any row vector $\mathbf{c}$ and any N-indexed sequence of GMM weighting matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots$ (which could be constant over N), the asymptotic variance $\operatorname*{plim}_{N\to\infty} N \| \mathbf{c}\widehat{\beta}_{\mathbf{A}_N} \|$ is smallest when $\operatorname*{plim}_{N\to\infty} \mathbf{A}_N = \mathbf{A}_{\mathrm{EGMM}}$.

We first show that $\operatorname*{plim}_{N\to\infty} N\left\langle \mathbf{c}\widehat{\beta}_{\mathbf{A}_N}, \mathbf{c}\widehat{\beta}_{\mathbf{A}_{\mathrm{EGMM}}} \right\rangle$ is invariant to the choice of sequence $(\mathbf{A}_N)$. We start with the definition of covariance and then use (6) and (7):

$$\begin{aligned}
&\operatorname*{plim}_{N\to\infty} N\langle \mathbf{c}\widehat{\beta}_{\mathbf{A}_N}, \mathbf{c}\widehat{\beta}_{\mathbf{A}_{\mathrm{EGMM}}}\rangle \\
&\quad= \operatorname*{plim}_{N\to\infty} N\,\mathrm{Cov}\Big\{ \mathbf{c}(\mathbf{S}'_{\mathbf{ZX}}\mathbf{A}_N\mathbf{S}_{\mathbf{ZX}})^{-1}\mathbf{S}'_{\mathbf{ZX}}\mathbf{A}_N\mathbf{S}_{\mathbf{ZY}}, \\
&\qquad\qquad \mathbf{c}(\mathbf{S}'_{\mathbf{ZX}}\mathbf{A}_{\mathrm{EGMM}}\mathbf{S}_{\mathbf{ZX}})^{-1}\mathbf{S}'_{\mathbf{ZX}}\mathbf{A}_{\mathrm{EGMM}}\mathbf{S}_{\mathbf{ZY}} \Big\} \\
&\quad= \left\{ \operatorname*{plim}_{N\to\infty} \mathbf{c}(\mathbf{\Sigma}'_{\mathbf{ZX}}\mathbf{A}_N\mathbf{\Sigma}_{\mathbf{ZX}})^{-1}\mathbf{\Sigma}'_{\mathbf{ZX}}\mathbf{A}_N N\,\mathrm{Var}(\mathbf{S_{ZY}}) \right\} \\
&\qquad\qquad \times \mathbf{A}_{\mathrm{EGMM}}\mathbf{\Sigma_{ZX}}(\mathbf{\Sigma}'_{\mathbf{ZX}}\mathbf{A}_{\mathrm{EGMM}}\mathbf{\Sigma_{ZX}})^{-1}\mathbf{c}' \\
&\quad= \left\{ \operatorname*{plim}_{N\to\infty} \mathbf{c}(\mathbf{\Sigma}'_{\mathbf{ZX}}\mathbf{A}_N\mathbf{\Sigma}_{\mathbf{ZX}})^{-1}\mathbf{\Sigma}'_{\mathbf{ZX}}\mathbf{A}_N\mathbf{A}^{-1}_{\mathrm{EGMM}} \right\} \\
&\qquad\qquad \times \mathbf{A}_{\mathrm{EGMM}}\mathbf{\Sigma_{ZX}}(\mathbf{\Sigma}'_{\mathbf{ZX}}\mathbf{A}_{\mathrm{EGMM}}\mathbf{\Sigma_{ZX}})^{-1}\mathbf{c}' \\
&\quad= \mathbf{c}\left\{ \operatorname*{plim}_{N\to\infty} (\mathbf{\Sigma}'_{\mathbf{ZX}}\mathbf{A}_N\mathbf{\Sigma}_{\mathbf{ZX}})^{-1}\mathbf{\Sigma}'_{\mathbf{ZX}}\mathbf{A}_N\mathbf{\Sigma}_{\mathbf{ZX}} \right\}(\mathbf{\Sigma}'_{\mathbf{ZX}}\mathbf{A}_{\mathrm{EGMM}}\mathbf{\Sigma_{ZX}})^{-1}\mathbf{c}' \\
&\quad= \mathbf{c}(\mathbf{\Sigma}'_{\mathbf{ZX}}\mathbf{A}_{\mathrm{EGMM}}\mathbf{\Sigma_{ZX}})^{-1}\mathbf{c}'
\end{aligned} \tag{8}$$

This does not depend on the sequence $(\mathbf{A}_N)$. As a result, for any $(\mathbf{A}_N)$,

$$\begin{aligned}\operatorname*{plim}_{N\to\infty} N\left\langle \mathbf{c}\widehat{\beta}_{\mathbf{A}_{\text{EGMM}}}, \mathbf{c}\left(\widehat{\beta}_{\mathbf{A}_{\text{EGMM}}} - \widehat{\beta}_{\mathbf{A}_N}\right)\right\rangle \\ = \operatorname*{plim}_{N\to\infty} N\left\langle \mathbf{c}\widehat{\beta}_{\mathbf{A}_{\text{EGMM}}}, \mathbf{c}\widehat{\beta}_{\mathbf{A}_{\text{EGMM}}}\right\rangle - \operatorname*{plim}_{N\to\infty} N\left\langle \mathbf{c}\widehat{\beta}_{\mathbf{A}_{\text{EGMM}}}, \mathbf{c}\widehat{\beta}_{\mathbf{A}_N}\right\rangle = 0\end{aligned}$$

That is, the difference between any linear GMM estimator and the EGMM estimator is asymptotically orthogonal to the latter. So by the Pythagorean theorem,

$$\operatorname*{plim}_{N\to\infty} N \parallel \mathbf{c}\widehat{\beta}_{\mathbf{A}_N} \parallel^2 = \operatorname*{plim}_{N\to\infty} N \parallel \mathbf{c}\widehat{\beta}_{\mathbf{A}_N} - \mathbf{c}\widehat{\beta}_{\mathbf{A}_{\text{EGMM}}} \parallel^2 + \operatorname*{plim}_{N\to\infty} N \parallel \mathbf{c}\widehat{\beta}_{\mathbf{A}_{\text{EGMM}}} \parallel^2$$

Thus $\operatorname*{plim}_{N\to\infty} N \parallel \mathbf{c}\widehat{\beta}_{\mathbf{A}_N} \parallel \geq \operatorname*{plim}_{N\to\infty} N \parallel \mathbf{c}\widehat{\beta}_{\mathbf{A}_{\text{EGMM}}} \parallel$. This suffices to prove the assertion that EGMM is, in fact, efficient. The result is akin to the fact that if there is a ball in midair, then the point on the ground closest to the ball (analogous to the efficient estimator) is such that the vector from the point to the ball is perpendicular to all vectors from the point to other spots on the ground, which are all inferior estimators of the ball's position.

Perhaps greater insight comes from a visualization based on another derivation of EGMM. Under the assumptions in our model, a direct OLS estimate of $\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}$ is biased. However, taking $\mathbf{Z}$-moments of both sides gives

$$\mathbf{Z}'\mathbf{Y} = \mathbf{Z}'\mathbf{X}\beta + \mathbf{Z}'\mathbf{E} \tag{9}$$

which is at least asymptotically amenable to OLS (Holtz-Eakin, Newey, and Rosen 1988). Still, OLS is not, in general, efficient on the transformed equation, because the error term, $\mathbf{Z}'\mathbf{E}$, is probably not independent and identically distributed (i.i.d.):
$\operatorname{Avar}\{(1/N)\mathbf{Z}'\mathbf{E}\} = \operatorname{Var}(\mathbf{z}\varepsilon)$, which cannot be assumed scalar. To solve this problem, we transform the equation again:

$$\operatorname{Var}(\mathbf{z}\varepsilon)^{-1/2}\mathbf{Z}'\mathbf{Y} = \operatorname{Var}(\mathbf{z}\varepsilon)^{-1/2}\mathbf{Z}'\mathbf{X}\beta + \operatorname{Var}(\mathbf{z}\varepsilon)^{-1/2}\mathbf{Z}'\mathbf{E}$$

Defining $\mathbf{X}^* = \operatorname{Var}(\mathbf{z}\varepsilon)^{-1/2}\mathbf{Z}'\mathbf{X}$, $\mathbf{Y}^* = \operatorname{Var}(\mathbf{z}\varepsilon)^{-1/2}\mathbf{Z}'\mathbf{Y}$, and $\mathbf{E}^* = \operatorname{Var}(\mathbf{z}\varepsilon)^{-1/2}\mathbf{Z}'\mathbf{E}$, the equation becomes

$$\mathbf{Y}^* = \mathbf{X}^*\beta + \mathbf{E}^* \tag{10}$$

By design now,

$$\begin{aligned}\operatorname{Avar}\left(\frac{1}{N}\mathbf{E}^*\right) &= \operatorname*{plim}_{N\to\infty} N\operatorname{Var}(\mathbf{z}\varepsilon)^{-1/2}\operatorname{Var}\left(\frac{1}{N}\mathbf{Z}'\mathbf{E}\right)\operatorname{Var}(\mathbf{z}\varepsilon)^{-1/2} \\ &= \operatorname{Var}(\mathbf{z}\varepsilon)^{-1/2}\operatorname{Var}(\mathbf{z}\varepsilon)\operatorname{Var}(\mathbf{z}\varepsilon)^{-1/2} = \mathbf{I}\end{aligned}$$

Because (10) has spherical errors, the Gauss–Markov theorem guarantees the efficiency of OLS applied to it, which is, by definition, generalized least squares (GLS) on (9): $\widehat{\beta}_{\text{GLS}} = \left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^*$. Unwinding with the definitions of $\mathbf{X}^*$ and $\mathbf{Y}^*$ yields EGMM, just as in (5).

Efficient GMM, then, is GLS on **Z**-moments. Where GLS projects **Y** into the column space of **X**, GMM estimators (efficient or otherwise) project $\mathbf{Z}'\mathbf{Y}$ into the column space of $\mathbf{Z}'\mathbf{X}$. These projections also map the variance ellipsoid of $\mathbf{Z}'\mathbf{Y}$, represented by $\text{Avar}\left\{(1/N)\mathbf{Z}'\mathbf{Y}\right\} = \text{Var}\left(\mathbf{z}\varepsilon\right)$, into the column space of $\mathbf{Z}'\mathbf{X}$. If $\text{Var}\left(\mathbf{z}\varepsilon\right)$ happens to be spherical, then the efficient projection is orthogonal, by Gauss–Markov, just as the shadow of a soccer ball is smallest when the sun is directly overhead. No reweighting of moments is needed for efficiency. But if the variance ellipsoid of the moments is an American football pointing at an odd angle, as in the examples at the beginning of this section—that is, if $\text{Var}\left(\mathbf{z}\varepsilon\right)$ is not scalar—then the efficient projection, the one casting the smallest shadow, is angled. To make that optimal projection, the mathematics in this second derivation stretch and shear space with a linear transformation to make the football spherical, perform an orthogonal projection, and then reverse the distortion.

2.3 Feasibility

Making EGMM practical requires a feasible estimator for the optimal weighting matrix, $\text{Var}\left(\mathbf{z}\varepsilon\right)^{-1}$. The usual route to this goal starts by observing that this matrix is the limit of an expression built around $\mathbf{\Omega}$:

$$\begin{aligned}
\text{Var}\left(\mathbf{z}\varepsilon\right) &= \operatorname*{plim}_{N\to\infty} N\,\text{Var}\left(\frac{1}{N}\mathbf{Z}'\mathbf{E}\right) = \operatorname*{plim}_{N\to\infty} NE\left(\frac{1}{N^2}\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z}\right) \\
&= \operatorname*{plim}_{N\to\infty} \frac{1}{N}E\left\{E\left(\mathbf{Z}'\mathbf{E}\mathbf{E}'\mathbf{Z}\,|\,\mathbf{Z}\right)\right\} \\
&= \operatorname*{plim}_{N\to\infty} \frac{1}{N}E\left\{\mathbf{Z}'E\left(\mathbf{E}\mathbf{E}\,|\,\mathbf{Z}\right)\mathbf{Z}\right\} = \operatorname*{plim}_{N\to\infty} \frac{1}{N}E\left(\mathbf{Z}'\mathbf{\Omega}\mathbf{Z}\right)
\end{aligned}$$

The simplest case is when the errors are believed to be homoskedastic, with $\mathbf{\Omega}$ of the form $\sigma^2\mathbf{I}$. Then, according to the last expression, the EGMM weighting matrix is the inverse of $\sigma^2 \operatorname*{plim}_{N\to\infty}(1/N)E\left(\mathbf{Z}'\mathbf{Z}\right)$, a consistent estimate of which is $(\sigma^2/N)\mathbf{Z}'\mathbf{Z}$. Plugging this choice for $\mathbf{A}$ into (2) and simplifying yields 2SLS:

$$\widehat{\beta}_{\text{2SLS}} = \left\{\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right\}^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y}$$

So when errors are i.i.d., 2SLS is EGMM.[3] When more-complex patterns of variance in the errors are suspected, the researcher can use a kernel-based estimator for the standard errors, such as the "sandwich" one ordinarily requested from Stata estimation commands with the `robust` and `cluster` options. A matrix, $\widehat{\mathbf{\Omega}}$, is constructed based on a formula that itself does not converge to $\mathbf{\Omega}$, but which has the property that $(1/N)\mathbf{Z}'\widehat{\mathbf{\Omega}}\mathbf{Z}$ is a consistent estimator of $\operatorname*{plim}_{N\to\infty}(1/N)E\left(\mathbf{Z}'\mathbf{\Omega}\mathbf{Z}\right)$ under given assumptions. $\left\{(1/N)\mathbf{Z}'\widehat{\mathbf{\Omega}}\mathbf{Z}\right\}^{-1}$

3. However, even when the two are asymptotically identical, in finite samples, the feasible EGMM algorithm we develop produces different results from 2SLS because it will, in practice, be based on a different moment weighting matrix.

or, equivalently, $\left(\mathbf{Z}'\widehat{\mathbf{\Omega}}\mathbf{Z}\right)^{-1}$ is then used as the weighting matrix. The result is the feasible EGMM (FEGMM) estimator:

$$\widehat{\beta}_{\text{FEGMM}} = \left\{\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\widehat{\mathbf{\Omega}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right\}^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\widehat{\mathbf{\Omega}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y} \tag{11}$$

For example, if we believe that the only deviation from sphericity is heteroskedasticity, then given consistent initial estimates, $\widehat{\mathbf{E}}$, of the residuals, we define

$$\widehat{\mathbf{\Omega}} = \begin{pmatrix} \widehat{e}_1^2 & & & \\ & \widehat{e}_2^2 & & \\ & & \ddots & \\ & & & \widehat{e}_N^2 \end{pmatrix}$$

Similarly, in a wide panel context, we can handle arbitrary patterns of covariance within individuals with a "clustered" $\widehat{\mathbf{\Omega}}$, a block-diagonal matrix with blocks

$$\widehat{\mathbf{\Omega}}_i = \widehat{\mathbf{E}}_i\widehat{\mathbf{E}}_i' = \begin{pmatrix} \widehat{e}_{i1}^2 & \widehat{e}_{i1}\widehat{e}_{i2} & \cdots & \widehat{e}_{i1}\widehat{e}_{iT} \\ \widehat{e}_{i2}\widehat{e}_{i1} & \widehat{e}_{2}^2 & \cdots & \widehat{e}_{i2}\widehat{e}_{iT} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{e}_{iT}\widehat{e}_{i1} & \cdots & \cdots & \widehat{e}_{iT}^2 \end{pmatrix} \tag{12}$$

Here $\widehat{\mathbf{E}}_i$ is the vector of residuals for individual i, the elements $\widehat{e}$ are double indexed for a panel, and T is the number of observations per individual.

A problem remains: where do the $\widehat{e}$ come from? They must be derived from an initial estimate of β. Fortunately, as long as the initial estimate is *consistent*, a GMM estimator fashioned from them is *efficient*: whatever valid algorithm is chosen to build the GMM weighting matrix will converge to the optimal matrix as N increases. Theoretically, any full-rank choice of $\mathbf{A}$ for the initial GMM estimate will suffice. Usual practice is to choose $\mathbf{A} = (\mathbf{Z}'\mathbf{H}\mathbf{Z})^{-1}$, where $\mathbf{H}$ is an "estimate" of $\mathbf{\Omega}$ based on a minimally arbitrary assumption about the errors, such as homoskedasticity.

Finally, we arrive at a practical recipe for linear GMM: perform an initial GMM regression, replacing $\widehat{\mathbf{\Omega}}$ in (11) with some reasonable but arbitrary $\mathbf{H}$, yielding $\widehat{\beta}_1$ (one-step GMM); obtain the residuals from this estimation; use the residuals to construct a sandwich proxy for $\mathbf{\Omega}$, calling it $\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}$; rerun the GMM estimation setting $\mathbf{A} = \left(\mathbf{Z}'\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{Z}\right)^{-1}$. This *two-step* estimator, $\widehat{\beta}_2$, is efficient and robust to whatever patterns of heteroskedasticity and cross correlation the sandwich covariance estimator models. In sum,

$$\begin{aligned} \widehat{\beta}_1 &= \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{H}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{H}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y} \\ \widehat{\beta}_2 &= \widehat{\beta}_{\text{FEGMM}} = \left\{\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right\}^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y} \end{aligned} \tag{13}$$

Historically, researchers often reported one-step results in addition to two-step results because of downward bias in the computed standard errors in two-step results. But as the next section explains, Windmeijer (2005) greatly reduces this problem.

2.4 Estimating standard errors

A derivation similar to that in (8) shows that the asymptotic variance of a linear GMM estimator is

$$\text{Avar}\left(\widehat{\beta}_{\mathbf{A}}\right) = \left(\mathbf{\Sigma}_{\mathbf{ZX}}'\mathbf{A}\mathbf{\Sigma}_{\mathbf{ZX}}\right)^{-1}\mathbf{\Sigma}_{\mathbf{ZX}}'\mathbf{A}\,\text{Var}\left(\mathbf{z}\varepsilon\right)\mathbf{A}\mathbf{\Sigma}_{\mathbf{ZX}}\left(\mathbf{\Sigma}_{\mathbf{ZX}}'\mathbf{A}\mathbf{\Sigma}_{\mathbf{ZX}}\right)^{-1} \tag{14}$$

But for both one- and two-step estimation, there are complications in developing feasible approximations for this formula.

In one-step estimation, although the choice of $\mathbf{A} = \left(\mathbf{Z}'\mathbf{H}\mathbf{Z}\right)^{-1}$ as a moment weighting matrix (discussed above) does not render the parameter estimates inconsistent even when $\mathbf{H}$ is based on incorrect assumptions about the variance of the errors, using $\mathbf{Z}'\mathbf{H}\mathbf{Z}$ to proxy for $\text{Var}\left(\mathbf{z}\varepsilon\right)$ in (14) would make the *variance* estimate for the parameters inconsistent. $\mathbf{Z}'\mathbf{H}\mathbf{Z}$ is not a consistent estimate of $\text{Var}\left(\mathbf{z}\varepsilon\right)$. The standard error estimates would not be "robust" to heteroskedasticity or serial correlation in the errors. Fortunately, they can be made so in the usual way, replacing $\text{Var}\left(\mathbf{z}\varepsilon\right)$ in (14) with a sandwich-type proxy based on the one-step residuals. This yields the feasible, robust estimator for the one-step standard errors:

$$\begin{aligned}\widehat{\text{Avar}}_{\text{r}}\left(\widehat{\beta}_1\right) = &\left\{\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{H}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right\}^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{H}\mathbf{Z}\right)^{-1}\mathbf{Z}'\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{H}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\\ &\times\left\{\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{H}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right\}^{-1}\end{aligned} \tag{15}$$

The complication with the two-step variance estimate is less straightforward. The thrust of the exposition to this point has been that, because of its sophisticated reweighting based on second moments, GMM is, in general, more efficient than 2SLS. But such assertions are asymptotic. Whether GMM is superior in finite samples—or whether the sophistication even backfires—is an empirical question. The case in point: for (infeasible) EGMM, in which $\mathbf{A} = \mathbf{A}_{\text{EGMM}} = \text{Var}\left(\mathbf{z}\varepsilon\right)^{-1}$, (14) simplifies to $\left\{\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{H}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right\}^{-1}$, a feasible, consistent estimate of which will typically be $\left\{\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right\}^{-1}$. This is the standard formula for the variance of linear GMM estimates. But it can produce standard errors that are downward biased when the number of instruments is large—severely enough to make two-step GMM useless for inference (Arellano and Bond 1991).

The trouble may be that in small samples, reweighting empirical moments based on their own estimated variances and covariances can end up mining data, indirectly overweighting observations that fit the model and underweighting ones that contradict it. The need to estimate the $j(j+1)/2$ distinct elements of the symmetric $\text{Var}\left(\mathbf{z}\varepsilon\right)$

can easily outstrip the statistical power of a small sample. These elements, as second moments of second moments, are fourth moments of the underlying data. When statistical power is that low, it becomes hard to distinguish moment means from moment variances—i.e., to distinguish third and fourth moments of the underlying data. For example, if the poorly estimated variance of some moment, $\text{Var}(\mathbf{z}_i\epsilon)$, is large, this could be because it truly has higher variance and deserves deemphasis; or it could be because the moment happens to put more weight on observations that do not fit the model well, in which case deemphasizing them overfits the model.

This phenomenon does not make coefficient estimates inconsistent because identification still flows from instruments believed to be exogenous. However, it can produce spurious precision in the form of implausibly small standard errors.

Windmeijer (2005) devised a small-sample correction for the two-step standard errors. The starting observation is that despite appearances in (13), $\widehat{\beta}_2$ is not simply linear in the random vector $\mathbf{Y}$. It is also a function of $\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}$, which depends on $\widehat{\beta}_1$, which depends on $\mathbf{Y}$ too. And the variance in $\mathbf{Y}$ is the ultimate source of the variance in the parameter estimates, through projection by the estimator. To express the full dependence of $\widehat{\beta}_2$ on $\mathbf{Y}$, let

$$g\left(\mathbf{Y},\widehat{\mathbf{\Omega}}\right)=\left\{\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\widehat{\mathbf{\Omega}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right\}^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\widehat{\mathbf{\Omega}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E} \tag{16}$$

By (3), this is the bias of the GMM estimator associated with $\mathbf{A}=\left(\mathbf{Z}'\widehat{\mathbf{\Omega}}\mathbf{Z}\right)^{-1}$. g is infeasible because the true disturbances, $\mathbf{E}$, are unobserved. In the second step of FEGMM, where $\widehat{\mathbf{\Omega}}=\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}$, $g\left(\mathbf{Y},\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\right)=\widehat{\beta}_2-\beta$, so g has the same variance as $\widehat{\beta}_2$, which is what we are interested in, but zero expectation.

Both of g's arguments are random. Yet the usual derivation of the variance estimate for $\widehat{\beta}_2$ treats $\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}$ as infinitely precise. That is appropriate for one-step GMM, where $\widehat{\mathbf{\Omega}}=\mathbf{H}$ is constant. But it is wrong in two-step GMM, where $\mathbf{Z}'\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{Z}$ is imprecise. To compensate, Windmeijer (2005) develops a formula for the dependence of g on the data via both of its arguments, and then calculates its variance. The expanded formula is infeasible, but a feasible approximation performs well in Windmeijer's simulations.

Windmeijer starts with a first-order Taylor expansion of g, viewed as a function of $\widehat{\beta}_1$, around the true (and unobserved) β:

$$g\left(\mathbf{Y},\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\right)\approx g\left(\mathbf{Y},\widehat{\mathbf{\Omega}}_{\beta}\right)+\frac{\partial}{\partial\widehat{\beta}}g\left(\mathbf{Y},\widehat{\mathbf{\Omega}}_{\widehat{\beta}}\right)\mid_{\widehat{\beta}=\beta}\left(\widehat{\beta}_1-\beta\right)$$

Defining $\mathbf{D}=\partial g\left(\mathbf{Y},\widehat{\mathbf{\Omega}}_{\widehat{\beta}}\right)/\partial\widehat{\beta}\mid_{\widehat{\beta}=\beta}$ and noting that $\widehat{\beta}_1-\beta=g\left(\mathbf{Y},\mathbf{H}\right)$, this is

$$g\left(\mathbf{Y},\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\right)\approx g\left(\mathbf{Y},\widehat{\mathbf{\Omega}}_{\beta}\right)+\mathbf{D}g\left(\mathbf{Y},\mathbf{H}\right) \tag{17}$$

Windmeijer expands the derivative in the definition of $\mathbf{D}$ using matrix calculus on (16), and then replaces infeasible terms within it, such as $\widehat{\mathbf{\Omega}}_{\beta}$, β, and $\mathbf{E}$, with feasible

approximations. It works out that the result, $\widehat{\mathbf{D}}$, is the $k \times k$ matrix whose pth column is

$$-\left\{\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right\}^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\frac{\partial\widehat{\mathbf{\Omega}}_{\widehat{\beta}}}{\partial\widehat{\beta}_p}\Big|_{\widehat{\beta}=\widehat{\beta}_1}\mathbf{Z}\left(\mathbf{Z}'\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\widehat{\mathbf{E}}_2$$

where $\widehat{\beta}_p$ is the pth element of $\widehat{\beta}$. The formula for the $\partial\widehat{\mathbf{\Omega}}_{\widehat{\beta}}/\partial\widehat{\beta}_p$ within this expression depends on that for $\widehat{\mathbf{\Omega}}_{\widehat{\beta}}$. For clustered errors on a panel, $\widehat{\mathbf{\Omega}}_{\widehat{\beta}}$ has blocks $\widehat{\mathbf{E}}_{1,i}\widehat{\mathbf{E}}'_{1,i}$, so by the product rule $\partial\widehat{\mathbf{\Omega}}_{\widehat{\beta}}/\partial\widehat{\beta}_p$, has blocks $\partial\widehat{\mathbf{E}}_{1,i}/\partial\widehat{\beta}_p\widehat{\mathbf{E}}'_{1,i} + \widehat{\mathbf{E}}_i\partial\widehat{\mathbf{E}}'_{1,i}/\partial\widehat{\beta}_p = -\mathbf{x}_{p,i}\widehat{\mathbf{E}}'_{1,i} - \widehat{\mathbf{E}}_{1,i}\mathbf{x}'_{p,i}$, where $\widehat{\mathbf{E}}_{1,i}$ contains the one-step errors for individual i and $\mathbf{x}_{p,i}$ holds the observations of regressor x_p for individual i. The feasible variance estimate of (17), i.e., the *corrected* estimate of the variance of $\widehat{\beta}_2$, works out to

$$\widehat{\text{Avar}}_{\text{c}}\left(\widehat{\beta}_2\right) = \widehat{\text{Avar}}\left(\widehat{\beta}_2\right) + \widehat{\mathbf{D}}\widehat{\text{Avar}}\left(\widehat{\beta}_2\right) + \widehat{\text{Avar}}\left(\widehat{\beta}_2\right)\widehat{\mathbf{D}}' + \widehat{\mathbf{D}}\widehat{\text{Avar}}_{\text{r}}\left(\widehat{\beta}_1\right)\widehat{\mathbf{D}}' \qquad (18)$$

The first term is the uncorrected variance estimate, and the last contains the robust one-step estimate. (The appendix provides a fuller derivation of the Windmeijer correction in the more general context of observation-weighted GMM.)

In difference GMM regressions on simulated panels, Windmeijer (2005) found that two-step EGMM performs somewhat better than one-step GMM in estimating coefficients, with lower bias and standard errors. And the reported two-step standard errors, with his correction, are quite accurate, so that two-step estimation with corrected errors seems modestly superior to cluster-robust one-step estimation.[4]

2.5 The Sargan/Hansen test of overidentifying restrictions

A crucial assumption for the validity of GMM is that the instruments are exogenous. If the model is exactly identified, detection of invalid instruments is impossible because even when $E(\mathbf{z}\varepsilon) \neq 0$, the estimator will choose $\widehat{\beta}$ so that $\mathbf{Z}'\widehat{\mathbf{E}} = \mathbf{0}$ exactly. But if the model is overidentified, a test statistic for the joint validity of the moment conditions (identifying restrictions) falls naturally out of the GMM framework. Under the null of joint validity, the vector of empirical moments, $(1/N)\mathbf{Z}'\widehat{\mathbf{E}}$, is randomly distributed around $\mathbf{0}$. A Wald test can check this hypothesis. If it holds, then the statistic

$$\left(\frac{1}{N}\mathbf{Z}'\widehat{\mathbf{E}}\right)'\text{Var}\left(\mathbf{z}\varepsilon\right)^{-1}\frac{1}{N}\mathbf{Z}'\widehat{\mathbf{E}} = \frac{1}{N}\left(\mathbf{Z}'\widehat{\mathbf{E}}\right)'\mathbf{A}_{\text{EGMM}}\mathbf{Z}'\widehat{\mathbf{E}}$$

is χ^2 with degrees of freedom equal to the degree of overidentification, $j - k$. The Hansen (1982) J test statistic for overidentifying restrictions is this expression made feasible by substituting a consistent estimate of $\mathbf{A}_{\text{EGMM}}$. It is just the minimized value of the criterion expression in (1) for a feasible EGMM estimator. If $\mathbf{\Omega}$ is scalar, then $\mathbf{A}_{\text{EGMM}} = (\mathbf{Z}'\mathbf{Z})^{-1}$. Here the Hansen test coincides with the Sargan (1958) test. But if

4. `xtabond2` offers both.

nonsphericity is suspected in the errors, as in robust one-step GMM, the Sargan statistic $(1/N)\left(\mathbf{Z}'\widehat{\mathbf{E}}\right)'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\widehat{\mathbf{E}}$ is inconsistent. Then a theoretically superior overidentification test for the *one-step* estimator is that based on the Hansen statistic from a *two-step* estimate. When the user requests the Sargan test for "robust" one-step GMM regressions, some software packages, including `ivreg2` and `xtabond2`, therefore quietly perform the second GMM step to obtain and report a consistent Hansen statistic.

Sargan/Hansen statistics can also be used to test the validity of subsets of instruments, via a "difference-in-Sargan/Hansen" test, also known as a C statistic. If one performs an estimation with and without a subset of suspect instruments, under the null of joint validity of the full instrument set, the difference in the two reported Sargan/Hansen test statistics is itself asymptotically χ^2, with degrees of freedom equal to the number of suspect instruments. The regression without the suspect instruments is called the "unrestricted" regression because it imposes fewer moment conditions. The difference-in-Sargan/Hansen test is, of course, only feasible if this unrestricted regression has enough instruments to be identified.

The Sargan/Hansen test should not be relied upon too faithfully, because it is prone to weakness. Intuitively speaking, when we apply it after GMM, we are first trying to drive $(1/N)\mathbf{Z}'\widehat{\mathbf{E}}$ close to 0, then testing whether it is close to 0. Counterintuitively, however, the test actually grows weaker the more moment conditions there are and, seemingly, the harder it should be to come close to satisfying them all.

2.6 The problem of too many instruments[5]

The difference and system GMM estimators described in the next section can generate moment conditions prolifically, with the instrument count quadratic in the time dimension of the panel, T. This can cause several problems in finite samples. First, because the number of elements in the estimated variance matrix of the moments is quadratic in the instrument count, it is *quartic* in T. A finite sample may lack adequate information to estimate such a large matrix well. It is not uncommon for the matrix to become singular, forcing the use of a generalized inverse.[6] This does not compromise consistency (again, any choice of $\mathbf{A}$ will give a consistent estimator), but does dramatize the distance of FEGMM from the asymptotic ideal. And it can weaken the Hansen test to the point where it generates implausibly good p-values of 1.000 (Anderson and Sørenson 1996; Bowsher 2002). Indeed, Sargan (1958) determined without the aid of modern computers that the error in his test is "proportional to the number of instrumental variables, so that, if the asymptotic approximations are to be used, this number must be small".

5. Roodman (2009) delves into the issues in this section.
6. `xtabond2` issues a warning when this happens.

A large instrument collection can overfit endogenous variables. For intuition, consider that in 2SLS, if the number of instruments equals the number of observations, the R^2s of the first-stage regressions are 1, and the second-stage results match those of (biased) OLS. This bias is present in all instrumental-variables (IV) regressions and becomes more pronounced as the instrument count rises.

Unfortunately, there appears to be little guidance from the literature on how many instruments is "too many" (Ruud 2000, 515), in part because the bias is present to some extent even when instruments are few. In one simulation of difference GMM on an 8×100 panel, Windmeijer (2005) reports that cutting the instrument count from 28 to 13 reduced the average bias in the two-step estimate of the parameter of interest by 40%. Simulations of panels of various dimensions in Roodman (2009) produce similar results. For instance, raising the instrument count from 5 to just 10 in a system GMM regression on a 5×100 panel raises the estimate of a parameter whose true value is 0.80 from 0.80 to 0.86. `xtabond2` warns when instruments outnumber individual units in the panel, as a minimally arbitrary rule of thumb; the simulations arguably indicate that that limit (equal to 100 here) is generous. At any rate, in using GMM estimators that can generate many instruments, it is good practice to report the instrument count and test the robustness of results to reducing it. The next sections describe the instrument sets typical of difference and system GMM, and ways to contain them with `xtabond2`.

3 The difference and system GMM estimators

The difference and system GMM estimators can be seen as part of a broader historical trend in econometric practice toward estimators that make fewer assumptions about the underlying data-generating process and use more complex techniques to isolate useful information. The plummeting costs of computation and software distribution no doubt have abetted the trend.

The difference and system GMM estimators are designed for panel analysis and embody the following assumptions about the data-generating process:

- The process may be dynamic, with current realizations of the dependent variable influenced by past ones.
- There may be arbitrarily distributed fixed individual effects. This argues against cross-section regressions, which must essentially assume fixed effects away, and in favor of a panel setup, where variation over time can be used to identify parameters.
- Some regressors may be endogenous.
- The idiosyncratic disturbances (those apart from the fixed effects) may have individual-specific patterns of heteroskedasticity and serial correlation.
- The idiosyncratic disturbances are uncorrelated across individuals.

Also, some secondary concerns shape the design:

- Some regressors can be predetermined but not strictly exogenous; that is, independent of current disturbances, some regressors can be influenced by past ones. The lagged dependent variable is an example.
- The number of time periods of available data, T, may be small. (The panel is "small T, large N".)

Finally, because the estimators are designed for general use, they do not assume that good instruments are available outside the immediate dataset. In effect, it is assumed that

- The only available instruments are "internal"—based on lags of the instrumented variables.

However, the estimators do allow inclusion of external instruments.

The general model of the data-generating process is much like that in section 2:

$$\begin{aligned} y_{it} &= \alpha y_{i,t-1} + \mathbf{x}'_{it}\beta + \varepsilon_{it} \\ \varepsilon_{it} &= \mu_i + v_{it} \\ E(\mu_i) &= E(v_{it}) = E(\mu_i v_{it}) = 0 \end{aligned} \tag{19}$$

Here the disturbance term has two orthogonal components: the fixed effects, μ_i, and the idiosyncratic shocks, v_{it}. We can rewrite (19) as

$$\Delta y_{it} = (\alpha - 1) y_{i,t-1} + \mathbf{x}'_{it}\beta + \varepsilon_{it} \tag{20}$$

So the model equally can be thought of as being for the level or increase of y.

In this section, we start with the classical OLS estimator applied to (19) and then modify it step by step to address all the concerns listed above, ending with the estimators of interest. For a continuing example, we will copy the application to firm-level employment in Arellano and Bond (1991). Their panel dataset is based on a sample of 140 UK firms surveyed annually from 1976 to 1984. The panel is unbalanced, with some firms having more observations than others. Because hiring and firing workers is costly, we expect employment to adjust with delay to changes in factors such as firms' capital stock, wages, and demand for the firms' output. The process of adjustment to changes in these factors may depend both on the passage of time, which indicates lagged versions of these factors as regressors, and on the difference between equilibrium employment level and the previous year's actual level, which argues for a dynamic model, in which lags of the dependent variable are also regressors.

The Arellano–Bond dataset can be downloaded with the Stata command `webuse abdata`.[7] The dataset indexes observations by the firm identifier, `id`, and `year`. The

7. In Stata 7, type `use http://www.stata-press.com/data/r7/abdata.dta`.

variable `n` is firm employment, `w` is the firm's wage level, `k` is the firm's gross capital, and `ys` is the aggregate output in the firm's sector, as a proxy for demand; all variables are in logarithms. Variable names ending in `L1` or `L2` indicate lagged copies. In their model, Arellano and Bond include the current and first lags of wages, the first two lags of employment, the current and first two lags of capital and sector-level output, and a set of time dummies.

A naïve attempt to estimate the model in Stata would look like this:

```
. webuse abdata
. regress n nL1 nL2 w wL1 k kL1 kL2 ys ysL1 ysL2 yr*
```

Source	SS	df	MS		
Model	1343.31797	16	83.9573732	Number of obs	= 751
Residual	7.57378164	734	.010318504	F(16, 734)	= 8136.58
				Prob > F	= 0.0000
Total	1350.89175	750	1.801189	R-squared	= 0.9944
				Adj R-squared	= 0.9943
				Root MSE	= .10158

n	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
nL1	1.044643	.0336647	31.03	0.000	.9785523	1.110734
nL2	-.0765426	.0328437	-2.33	0.020	-.1410214	-.0120639
w	-.5236727	.0487799	-10.74	0.000	-.6194374	-.427908
wL1	.4767538	.0486954	9.79	0.000	.381155	.5723527
k	.3433951	.0255185	13.46	0.000	.2932972	.3934931
kL1	-.2018991	.0400683	-5.04	0.000	-.2805613	-.123237
kL2	-.1156467	.0284922	-4.06	0.000	-.1715826	-.0597107
ys	.4328752	.1226806	3.53	0.000	.1920285	.673722
ysL1	-.7679125	.1658165	-4.63	0.000	-1.093444	-.4423813
ysL2	.3124721	.111457	2.80	0.005	.0936596	.5312846
yr1976	(dropped)					
yr1977	(dropped)					
yr1978	(dropped)					
yr1979	.0158888	.0143976	1.10	0.270	-.0123765	.0441541
yr1980	.0219933	.0166632	1.32	0.187	-.01072	.0547065
yr1981	-.0221532	.0204143	-1.09	0.278	-.0622306	.0179243
yr1982	-.0150344	.0206845	-0.73	0.468	-.0556422	.0255735
yr1983	.0073931	.0204243	0.36	0.717	-.0327038	.0474901
yr1984	.0153956	.0230101	0.67	0.504	-.0297779	.060569
_cons	.2747256	.3505305	0.78	0.433	-.4134363	.9628875

3.1 Purging fixed effects

One immediate problem in applying OLS to this empirical problem, and to (19) in general, is that $y_{i,t-1}$ is correlated with the fixed effects in the error term, which gives rise to "dynamic panel bias" (Nickell 1981). For example, consider the possibility that a firm experiences a large, negative employment shock for some reason not modeled, say, in 1980, so that the shock appears in the error term. All else equal, the apparent fixed effect for that firm for the entire 1976–1984 period—the deviation of its average unexplained employment from the sample average—will appear to be lower. In 1981, lagged employment and the fixed effect will *both* be lower. This positive correlation between a regressor and the error violates an assumption necessary for the consistency of OLS.

In particular, it inflates the coefficient estimate for lagged employment by attributing predictive power to it that actually belongs to the firm's fixed effect. Here $T = 9$. If T were large, the impact of one year's shock on the firm's apparent fixed effect would dwindle and so would the endogeneity problem.

There are two ways to work around this endogeneity. One, at the heart of difference GMM, is to transform the data to remove the fixed effects. The other is to instrument $y_{i,t-1}$ and any other similarly endogenous variables with variables thought uncorrelated with the fixed effects. System GMM incorporates that strategy and we will return to it.

An intuitive first attack on the fixed effects is to draw them out of the error term by entering dummies for each individual—the so-called least-squares dummy-variables (LSDV) estimator:

```
. xi: regress n nL1 nL2 w wL1 k kL1 kL2 ys ysL1 ysL2 yr* i.id
i.id              _Iid_1-140          (naturally coded; _Iid_1 omitted)
```

Source	SS	df	MS		
Model	1345.63898	155	8.68154179	Number of obs =	751
Residual	5.25277539	595	.008828194	F(155, 595) =	983.39
				Prob > F =	0.0000
Total	1350.89175	750	1.801189	R-squared =	0.9961
				Adj R-squared =	0.9951
				Root MSE =	.09396

n	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
nL1	.7329476	.039304	18.65	0.000	.6557563	.810139
nL2	-.1394773	.040026	-3.48	0.001	-.2180867	-.0608678
w	-.5597445	.057033	-9.81	0.000	-.6717551	-.4477339
wL1	.3149987	.0609756	5.17	0.000	.1952451	.4347522
k	.3884188	.0309544	12.55	0.000	.3276256	.4492119
kL1	-.0805185	.0384648	-2.09	0.037	-.1560618	-.0049751
kL2	-.0278013	.0328257	-0.85	0.397	-.0922695	.036667
ys	.468666	.1231278	3.81	0.000	.2268481	.7104839
ysL1	-.6285587	.15796	-3.98	0.000	-.9387856	-.3183318
ysL2	.0579764	.1345353	0.43	0.667	-.2062454	.3221982
yr1976	(dropped)					
yr1977	(dropped)					
yr1978	(dropped)					
yr1979	.0046562	.0137521	0.34	0.735	-.0223523	.0316647
yr1980	.0112327	.0164917	0.68	0.496	-.0211564	.0436218
yr1981	-.0253693	.0217036	-1.17	0.243	-.0679942	.0172557
yr1982	-.0343973	.0223548	-1.54	0.124	-.0783012	.0095066
yr1983	-.0280344	.0240741	-1.16	0.245	-.0753149	.0192461
yr1984	-.0119152	.0261724	-0.46	0.649	-.0633167	.0394862
_Iid_2	.2809286	.1197976	2.35	0.019	.0456511	.5162061
_Iid_3	.1147461	.0984317	1.17	0.244	-.0785697	.308062
(output omitted)						
_cons	1.821028	.495499	3.68	0.000	.8478883	2.794168

Or we could take advantage of another Stata command to do the same thing more succinctly:

```
. xtreg n nL1 nL2 w wL1 k kL1 kL2 ys ysL1 ysL2 yr*, fe
```

A third way to get nearly the same result is to partition the regression into two steps, first "partialling" the firm dummies out of the other variables with the Stata command `xtdata`, then running the final regression with those residuals. This partialling out applies a mean-deviations transform to each variable, where the mean is computed at the level of the firm. OLS on the data transformed is the within-groups estimator. It generates the same coefficient estimates, but it generates standard errors that are biased because they do not take into account the loss of N degrees of freedom in the pretransformation:[8]

```
. xtdata n nL1 nL2 w wL1 k kL1 kL2 ys ysL1 ysL2 yr*, fe
. regress n nL1 nL2 w wL1 k kL1 kL2 ys ysL1 ysL2 yr*
  (output omitted)
```

But within-groups does not eliminate dynamic panel bias (Nickell 1981; Bond 2002). Under the within-groups transformation, the lagged dependent variable becomes $y^*_{i,t-1} = y_{i,t-1} - \{1/(T-1)\}(y_{i2} + \cdots + y_{iT})$, while the error becomes $v^*_{it} = v_{it} - \{1/(T-1)\}(v_{i2} + \cdots + v_{iT})$. (The use of the lagged dependent variable as a regressor restricts the sample to $t = 2, \ldots, T$.) The problem is that the $y_{i,t-1}$ term in $y^*_{i,t-1}$ correlates negatively with the $-\{1/(T-1)\}v_{i,t-1}$ in v^*_{it}, while, symmetrically, the $-\{1/(T-1)\}y_{it}$ and v_{it} terms also move together.[9] So regressor and error are still correlated after transformation.

Worse, one cannot attack the continuing endogeneity by instrumenting $y^*_{i,t-1}$ with lags of $y_{i,t-1}$ (a strategy we will turn to soon) because they too are embedded in the transformed error, v^*_{it}. Again if T were large, then the $-\{1/(T-1)\}v_{i,t-1}$ and $-\{1/(T-1)\}y_{it}$ terms above would be insignificant, and the problem would disappear. But in simulations, Judson and Owen (1999) find a bias equal to 20% of the coefficient of interest even when $T = 30$.

Interestingly, where in our initial naïve OLS regression the lagged dependent variable was positively correlated with the error, biasing its coefficient estimate upward, the opposite is the case now. In the Stata examples, the estimate for the coefficient on lagged employment fell from 1.045 to 0.733. Good estimates of the true parameter should therefore lie in or near the range between these values. (In fact, a credible estimate should probably be below 1.00 because values above 1.00 imply an unstable dynamic, with accelerating divergence away from equilibrium values.) As Bond (2002) points out, these bounds provide a useful check on results from theoretically superior estimators.

Kiviet (1995) argues that the best way to handle dynamic panel bias is to perform LSDV, and then correct the results for the bias, which he finds can be predicted with surprising precision. However, the approach he advances works only for balanced panels and does not address the potential endogeneity of other regressors.

8. Because `xtdata` modifies the dataset, the dataset needs to be reloaded to copy later examples.
9. In fact, there are many other correlating term pairs, but their impact is second order because both terms in those pairs contain a $\{1/(T-1)\}$ factor.

As a result, the more practical strategy has been to develop estimators that theoretically need no correction. What is needed to directly remove dynamic panel bias is a different transformation of the data, one that expunges fixed effects while avoiding the propensity of the within-groups transformation to make every observation of y^* endogenous to every other for a given individual. There are many potential candidates. In fact, if the observations are sorted by individual within the data matrices $\mathbf{X}$ and $\mathbf{Y}$, then fixed effects can be purged by left-multiplying them by any block-diagonal matrix whose blocks each have width T and whose rows sum to zero. Such matrices map individual dummies to 0, thus purging fixed effects. How to choose? The transformation should have full row rank so that no further information is lost. It should make the transformed variables minimally dependent on lagged observations of the original variables so that they remain available as instruments. In other words, the blocks of the matrix should be upper triangular, or nearly so. A subtle, third criterion is that the transformation should be resilient to missing data—an idea we will clarify momentarily.

Two transformations are commonly used. One is the first-difference transform, which gives its name to "difference GMM". It is effected by $\mathbf{I}_N \otimes \mathbf{M}_\Delta$, where $\mathbf{I}_N$ is the identity matrix of order N and $\mathbf{M}_\Delta$ consists of a diagonal of -1s with a subdiagonal of 1s just to the right. Applying the transform to (19) gives

$$\Delta y_{it} = \alpha \Delta y_{i,t-1} + \Delta \mathbf{x}'_{it}\beta + \Delta v_{it}$$

Though the fixed effects are gone, the lagged dependent variable is still potentially endogenous, because the $y_{i,t-1}$ term in $\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2}$ is correlated with the $v_{i,t-1}$ in $\Delta v_{it} = v_{it} - v_{i,t-1}$. Likewise, any predetermined variables in $\mathbf{x}$ that are not strictly exogenous become potentially endogenous because they too may be related to $v_{i,t-1}$. But unlike with the mean-deviations transform, longer lags of the regressors remain orthogonal to the error and available as instruments.

The first-difference transform has a weakness. It magnifies gaps in unbalanced panels. If some y_{it} is missing, for example, then both Δy_{it} and $\Delta y_{i,t+1}$ are missing in the transformed data. One can construct datasets that completely disappear in first differences. This motivates the second common transformation, called "forward orthogonal deviations" or "orthogonal deviations" (Arellano and Bover 1995). Instead of subtracting the previous observation from the contemporaneous one, it subtracts the average of all future *available* observations of a variable. No matter how many gaps, it is computable for all observations except the last for each individual, so it minimizes data loss. And because lagged observations do not enter the formula, they are valid as instruments. To be precise, if w is a variable, then the transform is

$$w^{\perp}_{i,t+1} \equiv c_{it}\left(w_{it} - \frac{1}{T_{it}}\sum_{s>t} w_{is}\right) \tag{21}$$

where the sum is taken over available future observations, T_{it} is the number of such observations, and the scale factor, c_{it}, is $\sqrt{T_{it}/\left(T_{it}+1\right)}$. In a balanced panel, the transformation can be written cleanly as $\mathbf{I}_N \otimes \mathbf{M}_\perp$, where

$$\mathbf{M}_\perp = \begin{Bmatrix} \sqrt{\frac{T-1}{T}} & -\frac{1}{\sqrt{T(T-1)}} & -\frac{1}{\sqrt{T(T-1)}} & \cdots \\ & \sqrt{\frac{T-2}{T-1}} & -\frac{1}{\sqrt{(T-1)(T-2)}} & \cdots \\ & & \sqrt{\frac{T-3}{T-2}} & \cdots \\ & & & \ddots \end{Bmatrix}$$

One nice property of this transformation is that if the w_{it} are independently distributed before transformation, they remain so after. (The rows of $\mathbf{M}_\perp$ are orthogonal to each other.) The choice of c_{it} further assures that if the w_{it} are not only independent but also identically distributed, this property persists too; that is, $\mathbf{M}_\perp \mathbf{M}_\perp' = \mathbf{I}$.[10] This is not the case with differencing, which tends to make successive errors correlated even if they are uncorrelated before transformation: $\Delta v_{it} = v_{it} - v_{i,t-1}$ is mathematically related to $\Delta v_{i,t-1} = v_{i,t-1} - v_{i,t-2}$ via the shared $v_{i,t-1}$ term. However, researchers typically do not assume homoskedasticity in applying these estimators, so this property matters less than the resilience to gaps. In fact, Arellano and Bover show that in balanced panels, any two transformations of full row rank will yield numerically identical estimators, holding the instrument set fixed.

We will use the $*$ superscript to indicate data transformed by differencing or orthogonal deviations. The appearance of the $t+1$ subscript instead of t on the left side of (21) reflects the standard software practice of storing orthogonal deviations–transformed variables one period late, for consistency with the first-difference transform. With this definition, both transforms effectively drop the first observations for each individual; and for both, observations $w_{i,t-2}$ and earlier are the ones absent from the formula for w_{it}^*, making them valid instruments.

3.2 Instrumenting with lags

As emphasized at the beginning of this section, we are building an estimator for general application, in which we choose not to assume that the researcher has excellent instruments waiting in the wings. So we must draw instruments from within the dataset. Natural candidate instruments for $y_{i,t-1}^*$ are $y_{i,t-2}$ and, if the data are transformed by differencing, $\Delta y_{i,t-2}$. In the differenced case, for example, both $y_{i,t-2}$ and $\Delta y_{i,t-2}$ are mathematically related to $\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2}$ but not to the error term $\Delta v_{it} = v_{it} - v_{i,t-1}$ as long as the v_{it} are not serially correlated (see section 3.5). The simplest way to incorporate either instrument is with 2SLS, which leads us to the Anderson and Hsiao (1982) difference and levels estimators. Of these, the levels estimator, instrumenting with $y_{i,t-2}$ instead of $\Delta y_{i,t-2}$, seems preferable for maximizing sample size. $\Delta y_{i,t-2}$ is in general not available until $t = 4$, whereas $y_{i,t-2}$ is available

10. If $\mathrm{Var}\left(v_{it}\right) = \mathbf{I}$, then $\mathrm{Var}\left(\mathbf{M}_\perp v_{it}\right) = E\left(\mathbf{M}_\perp v_{it} v_{it}' \mathbf{M}_\perp'\right) = \mathbf{M}_\perp E\left(v_{it} v_{it}'\right) \mathbf{M}_\perp' = \mathbf{M}_\perp \mathbf{M}_\perp'$.

at $t = 3$, and an additional time period of data is significant in short panels. Returning to the employment example, we can implement the Anderson–Hsiao levels estimator (Anderson and Hsiao 1982) by using the Stata command `ivregress`:

```
. ivregress 2sls D.n (D.nL1= nL2) D.(nL2 w wL1 k kL1 kL2 ys ysL1 ysL2 yr1979
> yr1980 yr1981 yr1982 yr1983)
Instrumental variables (2SLS) regression              Number of obs =     611
                                                      Wald chi2(15) =   89.93
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =       .
                                                      Root MSE      =    .247
```

D.n	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
nL1						
D1.	2.307626	1.973193	1.17	0.242	-1.559762	6.175013
nL2						
D1.	-.2240271	.179043	-1.25	0.211	-.5749448	.1268907
w						
D1.	-.8103626	.261805	-3.10	0.002	-1.323491	-.2972342
wL1						
D1.	1.422246	1.179492	1.21	0.228	-.8895156	3.734007
k						
D1.	.2530975	.1447404	1.75	0.080	-.0305884	.5367835
kL1						
D1.	-.5524613	.6154929	-0.90	0.369	-1.758805	.6538825
kL2						
D1.	-.2126364	.2397909	-0.89	0.375	-.6826179	.2573451
ys						
D1.	.9905803	.4630105	2.14	0.032	.0830965	1.898064
ysL1						
D1.	-1.937912	1.438225	-1.35	0.178	-4.75678	.8809566
ysL2						
D1.	.4870838	.5099415	0.96	0.339	-.5123832	1.486551
yr1979						
D1.	.0467148	.0448599	1.04	0.298	-.0412089	.1346385
yr1980						
D1.	.0761344	.0624919	1.22	0.223	-.0463474	.1986163
yr1981						
D1.	.022623	.0557394	0.41	0.685	-.0866242	.1318701
yr1982						
D1.	.0127801	.0548402	0.23	0.816	-.0947048	.120265
yr1983						
D1.	.0099072	.0456113	0.22	0.828	-.0794894	.0993037
_cons	.0159337	.0273445	0.58	0.560	-.0376605	.0695279

```
Instrumented:  D.nL1
Instruments:   D.nL2 D.w D.wL1 D.k D.kL1 D.kL2 D.ys
               D.ysL1 D.ysL2 D.yr1979 D.yr1980 D.yr1981
               D.yr1982 D.yr1983 nL2
```

This is the first *consistent* estimate of the employment model, given our assumptions. It performs rather poorly, with a point estimate on the lagged dependent variable of 2.308, well outside the credible 0.733–1.045 range (between the LSDV and OLS point estimates discussed in section 3.1). The standard error on the coefficient is large too.

To improve efficiency, we can take the Anderson–Hsiao approach further, using longer lags of the dependent variable as additional instruments. To the extent this introduces more information, it should improve efficiency. But in standard 2SLS, the longer the lags used, the smaller the sample, because observations for which lagged observations are unavailable are dropped.

Working in the GMM framework, Holtz-Eakin, Newey, and Rosen (1988) show a way around this trade-off. As an example, standard 2SLS would enter the instrument $y_{i,t-2}$ into $\mathbf{Z}$ in one column, as a stack of blocks like

$$\mathbf{Z}_i = \begin{pmatrix} . \\ y_{i1} \\ \vdots \\ y_{i,T-2} \end{pmatrix}$$

The "." at the top represents a missing value, which forces the deletion of that row from the dataset. (Recall that the transformed variables being instrumented begin at $t = 2$, so the vector above starts at $t = 2$ and only its first observation lacks $y_{i,t-2}$.) Holtz-Eakin, Newey, and Rosen instead build a *set* of instruments from the second lag of y, one for each time period, and substitute zeros for missing observations, resulting in "GMM-style" instruments:

$$\begin{pmatrix} 0 & 0 & \cdots & 0 \\ y_{i1} & 0 & \cdots & 0 \\ 0 & y_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y_{i,T-2} \end{pmatrix}$$

In unbalanced panels, one also substitutes zeros for other missing values. These substitutions might seem like a dubious doctoring of the data in response to missing information. But the resulting columns of $\mathbf{Z}$, each taken as orthogonal to the transformed errors, correspond to a set of meaningful moment conditions,

$$E\left(\mathbf{Z}'\widehat{\mathbf{E}}\right) = \mathbf{0} \Rightarrow \sum_i y_{i,t-2}\widehat{e}^*_{it} = 0 \text{ for each } t \geq 3$$

which are based on an expectation we believe: $E\left(y_{i,t-2}\varepsilon^*_{it}\right) = 0$. (In fact, such instruments are perfectly valid, if unorthodox, in 2SLS, so "GMM-style" is a misleading label.) Alternatively, one could "collapse" this instrument set into one column:

$$\begin{pmatrix} 0 \\ y_{i1} \\ \vdots \\ y_{i,T-2} \end{pmatrix}$$

This embodies the same expectation but conveys slightly less information, because it generates one moment condition, $\sum_{i,t} y_{i,t-2}\widehat{e}^*_{it} = 0$.

Having eliminated the trade-off between lag length and sample length, it becomes practical to include *all* valid lags of the untransformed variables as instruments, where available. For endogenous variables, that means lags 2 and up. For a variable, w, that is predetermined but not strictly exogenous, lag 1 is also valid, because v^*_{it} is a function of errors no older than $v_{i,t-1}$ and $w_{i,t-1}$ is potentially correlated only with errors $v_{i,t-2}$ and older. For $y_{i,t-1}$, which is predetermined, realizations $y_{i,t-2}$ and earlier can be used, giving rise to stacked blocks in the instrument matrix of the form

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ y_{i1} & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & y_{i2} & y_{i1} & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & y_{i3} & y_{i2} & y_{i1} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \text{ or, collapsed, } \begin{pmatrix} 0 & 0 & 0 & \cdots \\ y_{i1} & 0 & 0 & \cdots \\ y_{i2} & y_{i1} & 0 & \cdots \\ y_{i3} & y_{i2} & y_{i1} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Because in the standard, uncollapsed form each instrumenting variable generates one column for each time period and lag available to that time period, the number of instruments is quadratic in T. To limit the instrument count (see section 2.6), one can restrict the lag ranges used in generating these instrument sets. Or one can collapse them. `xtabond2` allows both.[11]

Although these instrument sets are part of what defines difference (and system) GMM, researchers are free to incorporate other instruments. Given the importance of good instruments, it is worth giving serious thought to all options.

Returning to the employment example, the command line below expands on Anderson–Hsiao by generating GMM-style instruments for the lags of `n`, and then uses them in a 2SLS regression in differences. It treats all other regressors as exogenous; they instrument themselves, appearing in both the regressor matrix **X** and the instrument matrix **Z**. So **Z** contains both GMM-style instruments and ordinary one-column "IV-style" ones:

11. After conceiving of such instrument sets and adding a `collapse` option to `xtabond2`, I discovered precedents. Adapting Arellano and Bond's (1998) dynamic panel package, Dynamic Panel Data (DPD) for Gauss, and performing system GMM, Calderón, Chong, and Loayza (2002) use such instruments, followed by Beck and Levine (2004) and Moran, Graham, and Blomström (2005). Roodman (2009) demonstrates the superiority of collapsed instruments in some common situations with simulations.

```
. forvalues yr=1978/1984 {
  2.    forvalues lag = 2 / `= `yr´ - 1976´ {
  3.       quietly generate z`yr´L`lag´ = L`lag´.n if year == `yr´
  4.    }
  5. }
. quietly recode z* (. = 0)
. ivregress 2sls D.n D.(L2.n w L.w k L.k L2.k ys L.ys L2.ys yr1978 yr1979
> yr1980 yr1981 yr1982 yr1983) (DL.n = z*), nocons
note: z1978L2 dropped due to collinearity
note: z1984L3 dropped due to collinearity
Instrumental variables (2SLS) regression              Number of obs =     611
                                                      Wald chi2(16) =       .
                                                      Prob > chi2   =       .
                                                      R-squared     =       .
                                                      Root MSE      =  .10885
```

D.n	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
n						
LD.	.2689418	.1447008	1.86	0.063	-.0146665	.5525501
L2D.	-.0669834	.0431623	-1.55	0.121	-.1515799	.0176131
w						
D1.	-.5723355	.0573518	-9.98	0.000	-.684743	-.459928
LD.	.2112242	.1037099	2.04	0.042	.0079564	.4144919
k						
D1.	.3843826	.0319335	12.04	0.000	.3217941	.4469711
LD.	.0796079	.0538637	1.48	0.139	-.025963	.1851788
L2D.	.0231674	.0364836	0.64	0.525	-.0483391	.094674
ys						
D1.	.5976429	.119675	4.99	0.000	.3630842	.8322016
LD.	-.4806272	.1614112	-2.98	0.003	-.7969874	-.164267
L2D.	.0581721	.1340154	0.43	0.664	-.2044932	.3208374
yr1978						
D1.	.0429548	.0427985	1.00	0.316	-.0409287	.1268384
yr1979						
D1.	.047082	.0410286	1.15	0.251	-.0333325	.1274966
yr1980						
D1.	.0566061	.0390769	1.45	0.147	-.0199833	.1331955
yr1981						
D1.	.0263295	.0357589	0.74	0.462	-.0437567	.0964158
yr1982						
D1.	.0018456	.0280028	0.07	0.947	-.0530388	.0567301
yr1983						
D1.	-.0062288	.0195166	-0.32	0.750	-.0444805	.032023

```
Instrumented:  LD.n
Instruments:   L2D.n D.w LD.w D.k LD.k L2D.k D.ys LD.ys
               L2D.ys D.yr1978 D.yr1979 D.yr1980 D.yr1981
               D.yr1982 D.yr1983 z1979L2 z1979L3 z1980L2
               z1980L3 z1980L4 z1981L2 z1981L3 z1981L4
               z1981L5 z1982L2 z1982L3 z1982L4 z1982L5
               z1982L6 z1983L2 z1983L3 z1983L4 z1983L5
               z1983L6 z1983L7 z1984L2 z1984L4 z1984L5
               z1984L6 z1984L7 z1984L8
```

Although this estimate is, in theory, not only consistent but also more efficient than Anderson–Hsiao, it still seems poorly behaved. Now the coefficient estimate for lagged employment has plunged to 0.269, about 3 standard errors below the 0.733–1.045 range.

What is going on? As discussed in section 2.2, 2SLS is efficient under homoskedasticity. But after differencing, the disturbances Δv_{it} may be far from independent, apparently far enough to greatly reduce accuracy. $\Delta v_{it} = v_{it} - v_{i,t-1}$ can be correlated with $\Delta v_{i,t-1} = v_{i,t-1} - v_{i,t-2}$, with which it shares a $v_{i,t-1}$ term. Feasible GMM directly addresses this problem, modeling the error structure more realistically, which makes it both more precise asymptotically and better behaved in practice.[12]

3.3 Applying GMM

The only way errors could reasonably be expected to be spherical in "difference GMM" is if a) the untransformed errors are i.i.d., which is usually not assumed, and b) the orthogonal deviations transform is used, so that the errors remain spherical. Otherwise, as section 2.2 showed, FEGMM is asymptotically superior.

To implement FEGMM, however, we must estimate $\mathbf{\Omega}^*$, the covariance matrix of the transformed errors—and do so twice for two-step GMM. For the first step, the least arbitrary choice of $\mathbf{H}$, the a priori estimate of $\mathbf{\Omega}^*$ (see section 2.3), is based, ironically, on the assumption that the v_{it} are i.i.d. after all. Using this assumption, and letting $\mathbf{v}_i$ refer to the vector of idiosyncratic errors for individual i, we set $\mathbf{H}$ to $\mathbf{I}_N \otimes \operatorname{Var}(\mathbf{v}_i^* \mid \mathbf{Z})$, where

$$\operatorname{Var}(\mathbf{v}_i^* \mid \mathbf{Z}) = \operatorname{Var}(\mathbf{M}_* \mathbf{v}_i \mid \mathbf{Z}) = \mathbf{M}_* \operatorname{Var}(\mathbf{v}_i \mathbf{v}_i' \mid \mathbf{Z}) \mathbf{M}_*' = \mathbf{M}_* \mathbf{M}_*' \tag{22}$$

For orthogonal deviations, this is $\mathbf{I}$, as discussed in section 3.1. For differences, it is

$$\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & \ddots \\ & & \ddots & \ddots \end{pmatrix} \tag{23}$$

As for the second FEGMM step, here we proxy $\mathbf{\Omega}^*$ with the robust, clustered estimate in (12), which is built on the assumption that errors are correlated only within individuals, not across them. For this reason, it is almost always wise to include time dummies to remove universal time-related shocks from the errors.

With these choices, we reach the classic Arellano and Bond (1991) difference GMM estimator for dynamic panels. As the name suggests, Arellano and Bond originally proposed using the differencing transform. When orthogonal deviations are used instead, perhaps the estimator ought to be called "deviations GMM"—but the term is not common.

Pending the full definition of the `xtabond2` syntax in section 4.1, the example in this section shows how to use the command to estimate the employment equation from before. First, the final estimates in the previous section can actually be obtained from `xtabond2` by typing

12. Apparent bias toward 0 in the coefficient estimate can also indicate weak instrumentation, a concern that motivates "system GMM," discussed later.

```
. xtabond2 n L.n L2.n w L.w L(0/2).(k ys) yr*, gmm(L.n)
> iv(w L.w L(0/2).(k ys) yr*) h(1) nolevel small
```

The `h(1)` option here specifies $\mathbf{H} = \mathbf{I}$, which embodies the incorrect assumption of homoskedasticity. If we drop that, $\mathbf{H}$ defaults to the form given in (23), and the results greatly improve:

```
. xtabond2 n L.n L2.n w L.w L(0/2).(k ys) yr*, gmm(L.n)
> iv(w L.w L(0/2).(k ys) yr*) nolevel robust
Favoring space over speed. To switch, type or click on mata: mata set matafavor
> speed, perm.
yr1976 dropped due to collinearity
yr1977 dropped due to collinearity
yr1984 dropped due to collinearity
Warning: Two-step estimated covariance matrix of moments is singular.
  Using a generalized inverse to calculate robust weighting matrix for Hansen
> test.
  Difference-in-Sargan statistics may be negative.
Dynamic panel-data estimation, one-step difference GMM
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       611
Time variable : year                            Number of groups   =       140
Number of instruments = 41                      Obs per group: min =         4
Wald chi2(16) =   1727.45                                      avg =      4.36
Prob > chi2   =     0.000                                      max =         6
------------------------------------------------------------------------------
             |               Robust
           n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           n |
          L1.|   .6862261   .1445943     4.75   0.000     .4028266    .9696257
          L2.|  -.0853582   .0560155    -1.52   0.128    -.1951467    .0244302
             |
           w |
          --.|  -.6078208   .1782055    -3.41   0.001    -.9570972   -.2585445
          L1.|   .3926237   .1679931     2.34   0.019     .0633632    .7218842
             |
           k |
          --.|   .3568456   .0590203     6.05   0.000      .241168    .4725233
          L1.|  -.0580012   .0731797    -0.79   0.428    -.2014308    .0854284
          L2.|  -.0199475   .0327126    -0.61   0.542    -.0840631    .0441681
             |
          ys |
          --.|   .6085073   .1725313     3.53   0.000     .2703522    .9466624
          L1.|  -.7111651   .2317163    -3.07   0.002    -1.165321   -.2570095
          L2.|   .1057969   .1412021     0.75   0.454    -.1709542     .382548
             |
      yr1978 |   .0077033   .0314106     0.25   0.806    -.0538604     .069267
      yr1979 |   .0172578   .0290922     0.59   0.553    -.0397619    .0742775
      yr1980 |   .0297185   .0276617     1.07   0.283    -.0244974    .0839344
      yr1981 |   -.004071   .0298987    -0.14   0.892    -.0626713    .0545293
      yr1982 |  -.0193555   .0228436    -0.85   0.397     -.064128    .0254171
      yr1983 |  -.0136171   .0188263    -0.72   0.469     -.050516    .0232818
------------------------------------------------------------------------------
```

(*Continued on next page*)

```
Instruments for first differences equation
  Standard
    D.(w L.w k L.k L2.k ys L.ys L2.ys yr1976 yr1977 yr1978 yr1979 yr1980
    yr1981 yr1982 yr1983 yr1984)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/.).L.n
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -3.60  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.52  Pr > z =  0.606
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(25)   =  67.59  Prob > chi2 =  0.000
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(25)   =  31.38  Prob > chi2 =  0.177
  (Robust, but can be weakened by many instruments.)
Difference-in-Hansen tests of exogeneity of instrument subsets:
  iv(w L.w k L.k L2.k ys L.ys L2.ys yr1976 yr1977 yr1978 yr1979 yr1980 yr1981
> yr1982 yr1983 yr1984)
    Hansen test excluding group:     chi2(11)   =  12.01  Prob > chi2 =  0.363
    Difference (null H = exogenous): chi2(14)   =  19.37  Prob > chi2 =  0.151
```

To obtain two-step estimates, we would merely change `robust` to `twostep`. These commands exactly match the one- and two-step results in Arellano and Bond (1991).[13] Even so, the one-step coefficient on lagged employment of 0.686 (and the two-step coefficient of 0.629) is not quite in the hoped for 0.733–1.045 range, which hints at specification problems. Interestingly, Blundell and Bond (1998) write that they "do not expect wages and capital to be strictly exogenous in our employment application", but the above regressions assume just that. If we instrument them too, in GMM style, then the coefficient on lagged employment moves into the credible range:

```
. xtabond2 n L.n L2.n w L.w L(0/2).(k ys) yr*, gmm(L.(n w k)) iv(L(0/2).ys yr*)
> nolevel robust small
Favoring space over speed. To switch, type or click on mata: mata set matafavor
> speed, perm.
yr1976 dropped due to collinearity
yr1977 dropped due to collinearity
yr1984 dropped due to collinearity
Warning: Two-step estimated covariance matrix of moments is singular.
  Using a generalized inverse to calculate robust weighting matrix for Hansen
> test.
  Difference-in-Sargan statistics may be negative.
Dynamic panel-data estimation, one-step difference GMM
------------------------------------------------------------------------------
```

13. See table 4, columns (a1) and (a2) in that article.

```
Group variable: id                              Number of obs      =       611
Time variable : year                            Number of groups   =       140
Number of instruments = 90                      Obs per group: min =         4
F(16, 140)    =     85.30                                      avg =      4.36
Prob > F      =     0.000                                      max =         6
```

n	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
n						
L1.	.8179867	.0859761	9.51	0.000	.6480073	.987966
L2.	-.1122756	.0502366	-2.23	0.027	-.211596	-.0129552
w						
--.	-.6816685	.1425813	-4.78	0.000	-.9635594	-.3997776
L1.	.6557083	.202368	3.24	0.001	.2556158	1.055801
k						
--.	.3525689	.1217997	2.89	0.004	.1117643	.5933735
L1.	-.1536626	.0862928	-1.78	0.077	-.324268	.0169428
L2.	-.0304529	.0321355	-0.95	0.345	-.0939866	.0330807
ys						
--.	.6509498	.189582	3.43	0.001	.276136	1.025764
L1.	-.9162028	.2639274	-3.47	0.001	-1.438001	-.3944042
L2.	.2786584	.1855286	1.50	0.135	-.0881415	.6454584
yr1978	.0238987	.0367972	0.65	0.517	-.0488513	.0966487
yr1979	.0352258	.0351846	1.00	0.318	-.034336	.1047876
yr1980	.0502675	.0365659	1.37	0.171	-.0220252	.1225602
yr1981	.0102721	.0349996	0.29	0.770	-.058924	.0794683
yr1982	-.0111623	.0264747	-0.42	0.674	-.0635042	.0411797
yr1983	-.0069458	.0191611	-0.36	0.718	-.0448283	.0309368

```
Instruments for first differences equation
  Standard
    D.(ys L.ys L2.ys yr1976 yr1977 yr1978 yr1979 yr1980 yr1981 yr1982 yr1983
    yr1984)
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/.).(L.n L.w L.k)
------------------------------------------------------------------------------
Arellano-Bond test for AR(1) in first differences: z =  -5.39  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.78  Pr > z =  0.436
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(74)   = 120.62  Prob > chi2 =  0.001
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(74)   =  73.72  Prob > chi2 =  0.487
  (Robust, but can be weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  iv(ys L.ys L2.ys yr1976 yr1977 yr1978 yr1979 yr1980 yr1981 yr1982 yr1983
> yr1984)
    Hansen test excluding group:     chi2(65)   =  56.99  Prob > chi2 =  0.750
    Difference (null H = exogenous): chi2(9)    =  16.72  Prob > chi2 =  0.053
```

3.4 Instrumenting with variables orthogonal to the fixed effects

Arellano and Bond compare the performance of one- and two-step difference GMM with the OLS, within-groups, and Anderson–Hsiao difference and levels estimators using Monte Carlo simulations of 7×100 panels. Difference GMM exhibits the least bias and

variance in estimating the parameter of interest, although in their tests the Anderson–Hsiao levels estimator does nearly as well for most parameter choices. But there are many degrees of freedom in designing such tests. As Blundell and Bond (1998) demonstrate in separate simulations, if y is close to a random walk, then difference GMM performs poorly because past levels convey little information about future changes, so untransformed lags are weak instruments for transformed variables.

To increase efficiency, under an additional assumption, Blundell and Bond develop an approach outlined in Arellano and Bover (1995), pursuing the second strategy against dynamic panel bias offered in section 3.1. Instead of transforming the regressors to expunge the fixed effects, it transforms—differences—the instruments to make them exogenous to the fixed effects. This is valid assuming that changes in any instrumenting variable, w, are uncorrelated with the fixed effects: $E(\Delta w_{it}\mu_i) = 0$ for all i and t. This is to say, $E(w_{it}\mu_i)$ is time-invariant. If this holds, then $\Delta w_{i,t-1}$ is a valid instrument for the variables in levels:

$$E(\Delta w_{i,t-1}\varepsilon_{it}) = E(\Delta w_{i,t-1}\mu_i) + E(w_{i,t-1}v_{it}) - E(w_{i,t-2}v_{it}) = 0 + 0 - 0$$

In a nutshell, where Arellano–Bond instruments differences (or orthogonal deviations) with levels, Blundell–Bond instruments levels with differences. For random walk–like variables, past changes may indeed be more predictive of current levels than past levels are of current changes so that the new instruments are more relevant. Again validity depends on the assumption that the v_{it} are not serially correlated. Otherwise, $w_{i,t-1}$ and $w_{i,t-2}$, correlated with past and contemporary errors, may correlate with future ones as well. In general, if w is endogenous, $\Delta w_{i,t-1}$ is available as an instrument because $\Delta w_{i,t-1} = w_{i,t-1} - w_{i,t-2}$ should not correlate with v_{it}; earlier realizations of Δw can serve as instruments as well. And if w is predetermined, the contemporaneous $\Delta w_{it} = w_{it} - w_{i,t-1}$ is also valid, because $E(w_{it}v_{it}) = 0$.

But the new assumption is not trivial; it is akin to one of stationarity. The Blundell–Bond approach instruments $y_{i,t-1}$ with $\Delta y_{i,t-1}$, which from the point of view of (20) contains the fixed effect μ_i—yet we assume that the levels equation error, ε_{it}, contains μ_i too, which makes the proposition that the instrument is orthogonal to the error, that $E(\Delta y_{i,t-1}\varepsilon_{it}) = 0$, counterintuitive. The assumption can hold, but only if the data-generating process is such that the fixed effect and the autoregressive process governed by α, the coefficient on the lagged dependent variable, offset each other in expectation across the whole panel, much like investment and depreciation in a Solow growth model steady state.

Blundell and Bond formalize this idea.[14] They stipulate that α must have absolute value less than unity so that the process converges. Then they derive the assumption that $E(\Delta w_{it}\mu_i) = 0$ from a more precise one about the initial conditions of the data-generating process. It is easiest to state for the simple autoregressive model without controls: $y_{it} = \alpha y_{i,t-1} + \mu_i + v_{it}$. Conditioning on μ_i, y_{it} can be expected to converge over time to $\mu_i/(1-\alpha)$—the point where the fixed effect and the autoregressive decay just offset each other.[15] For the time-invariance of $E(y_{it}\mu_i)$ to hold, the deviations of

14. Roodman (2009) provides a pedagogic introduction to these ideas.
15. This can be seen by solving $E(y_{it} \mid \mu_i) = E(y_{i,t-1} \mid \mu_i)$, using $y_{it} = \alpha y_{i,t-1} + \mu_i + v_{it}$.

the initial observations, y_{i1}, from these long-term convergent values must not correlate with the fixed effects: $E\left[\mu_i \left\{y_{i1} - \mu_i / (1-\alpha)\right\}\right] = 0$. Otherwise, the "regression to the mean" that will occur, whereby individuals with higher initial deviations will have slower subsequent changes as they converge to the long-run mean, will correlate with the fixed effects in the error. If this condition is satisfied in the first period, then it will be in subsequent ones as well. Generalizing to models with controls $\mathbf{x}$, this assumption about initial conditions is that, controlling for the covariates, faster-growing individuals are not systematically closer or farther from their steady states than slower-growing ones.

To exploit the new moment conditions for the data in levels while retaining the original Arellano–Bond conditions for the transformed equation, Blundell and Bond designed a *system* estimator. This involved building a stacked dataset with twice the observations; in each individual's data, the untransformed observations follow the transformed ones. Formally, we produce the augmented, transformed dataset by left-multiplying the original by an augmented transformation matrix,

$$\mathbf{M}_*^+ = \begin{pmatrix} \mathbf{M}_* \\ \mathbf{I} \end{pmatrix}$$

where $\mathbf{M}_* = \mathbf{M}_\Delta$ or $\mathbf{M}_\perp$. Thus, for individual i, the augmented dataset is

$$\mathbf{X}_i^+ = \begin{pmatrix} \mathbf{X}_i * \\ \mathbf{X}_i \end{pmatrix}, \mathbf{Y}_i^+ = \begin{pmatrix} \mathbf{Y}_i^* \\ \mathbf{Y}_i \end{pmatrix}$$

The GMM formulas and the software treat the system as a single-equation estimation problem because the same linear relationship with the same coefficients is believed to apply to both the transformed and untransformed variables.

In system GMM, one can include time-invariant regressors, which would disappear in difference GMM. Asymptotically, this does not affect the coefficient estimates for other regressors because all instruments for the levels equation are assumed to be orthogonal to fixed effects, indeed to all time-invariant variables. In expectation, removing them from the error term does not affect the moments that are the basis for identification. However, it is still a mistake to introduce explicit fixed-effects dummies, for they would still effectively cause the within-groups transformation to be applied as described in section 3.1. In fact, any dummy that is 0 for almost all individuals, or 1 for almost all, might cause bias in the same way, especially if T is very small.

The construction of the augmented instrument matrix, $\mathbf{Z}^+$, is somewhat more complicated. For a one-column, IV-style instrument, a strictly exogenous variable, w, with observation vector $\mathbf{W}$, could be transformed and entered like the regressors above,

$$\begin{pmatrix} \mathbf{W}^* \\ \mathbf{W} \end{pmatrix} \tag{24}$$

imposing the moment condition $\sum w_{it}^* \widehat{e}_{it}^* + \sum w_{it} \widehat{e}_{it} = 0$. Alternative arrangements, implying slightly different conditions, include

$$\begin{pmatrix} \mathbf{0} \\ \mathbf{W} \end{pmatrix} \text{ and } \begin{pmatrix} \mathbf{W}^* & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{pmatrix} \tag{25}$$

As for GMM-style instruments, the Arellano–Bond instruments for the transformed data are set to zero for levels observations, and the new instruments for the levels data are set to zero for the transformed observations. One could enter a full GMM-style set of differenced instruments for the levels equation, using all available lags, in direct analogy with the levels instruments entered for the transformed equation. However, most of these would be mathematically redundant in system GMM. The figure below shows why, with the example of a predetermined variable, w, under the difference transform.[16] The $\underline{\underline{D}}$ symbols link moments equated by the Arellano–Bond conditions on the differenced equation. The upper-left one, for example, asserts that $E(w_{i1}\varepsilon_{i2}) = E(w_{i1}\varepsilon_{i1})$, which is equivalent to the Arellano–Bond moment condition, $E(w_{i1}\Delta\varepsilon_{i2}) = 0$. The $\| L$ symbols do the same for the new Arellano–Bover conditions.

$$
\begin{array}{cccccccc}
E(w_{i1}\varepsilon_{i1}) & \underline{\underline{D}} & E(w_{i1}\varepsilon_{i2}) & \underline{\underline{D}} & E(w_{i1}\varepsilon_{i3}) & \underline{\underline{D}} & E(w_{i1}\varepsilon_{i4}) \\
 & & \| L & & & & \\
E(w_{i2}\varepsilon_{i1}) & & E(w_{i2}\varepsilon_{i2}) & \underline{\underline{D}} & E(w_{i2}\varepsilon_{i3}) & \underline{\underline{D}} & E(w_{i2}\varepsilon_{i4}) \\
 & & & & \| L & & \\
E(w_{i3}\varepsilon_{i1}) & & E(w_{i3}\varepsilon_{i2}) & & E(w_{i3}\varepsilon_{i3}) & \underline{\underline{D}} & E(w_{i3}\varepsilon_{i4}) \\
 & & & & & & \| L \\
E(w_{i4}\varepsilon_{i1}) & & E(w_{i4}\varepsilon_{i2}) & & E(w_{i4}\varepsilon_{i3}) & & E(w_{i4}\varepsilon_{i4})
\end{array}
$$

One could add more vertical links to the upper triangle of the grid, but it would add no new information. The ones included above embody the moment restrictions $\sum_i \Delta w_{it}\varepsilon_{it} = 0$ for each $t > 1$. If w is endogenous, those conditions become invalid because the w_{it} in Δw_{it} is endogenous to the v_{it} in ε_{it}. Lagging w one period sidesteps this endogeneity, yielding the valid moment conditions $\sum_i \Delta w_{i,t-1}\varepsilon_{it} = 0$ for each $t > 2$:

$$
\begin{array}{ccccccc}
E(w_{i1}\varepsilon_{i1}) & E(w_{i1}\varepsilon_{i2}) & \underline{\underline{D}} & E(w_{i1}\varepsilon_{i3}) & \underline{\underline{D}} & E(w_{i1}\varepsilon_{i4}) \\
 & & & \| L & & \\
E(w_{i2}\varepsilon_{i1}) & E(w_{i2}\varepsilon_{i2}) & & E(w_{i2}\varepsilon_{i3}) & \underline{\underline{D}} & E(w_{i2}\varepsilon_{i4}) \\
 & & & & & \| L \\
E(w_{i3}\varepsilon_{i1}) & E(w_{i3}\varepsilon_{i2}) & & E(w_{i3}\varepsilon_{i3}) & & E(w_{i3}\varepsilon_{i4}) \\
 & & & & & \\
E(w_{i4}\varepsilon_{i1}) & E(w_{i4}\varepsilon_{i2}) & & E(w_{i4}\varepsilon_{i3}) & & E(w_{i4}\varepsilon_{i4})
\end{array}
$$

If w is predetermined, the new moment conditions translate into the system GMM instrument matrix with blocks of the form

$$
\begin{pmatrix}
0 & 0 & 0 & 0 & \cdots \\
\triangle w_{i2} & 0 & 0 & 0 & \cdots \\
0 & \triangle w_{i3} & 0 & 0 & \cdots \\
0 & 0 & \triangle w_{i4} & 0 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix}
\text{ or , collapsed, }
\begin{pmatrix}
0 \\ \triangle w_{i2} \\ \triangle w_{i3} \\ \triangle w_{i4} \\ \vdots
\end{pmatrix}
$$

Here the first row of the matrix corresponds to $t = 1$. If w is endogenous, then the nonzero elements are shifted down one row.

16. Tue Gørgens devised these diagrams.

Again the last item of business is defining $\mathbf{H}$, which now must be seen as a preliminary variance estimate for the augmented error vector, $\mathbf{E}^+$. As before, to minimize arbitrariness we set $\mathbf{H}$ to what $\text{Var}(\mathbf{E}^+)$ would be in the simplest case. This time, however, assuming homoskedasticity and unit variance of the idiosyncratic errors does not suffice to define a unique $\mathbf{H}$, because the fixed effects are present in the levels errors. Consider, for example, $\text{Var}(\varepsilon_{it})$, for some i, t, which is on the diagonal of $\text{Var}(\mathbf{E}^+)$. Expanding this, we have

$$\text{Var}(\varepsilon_{it}) = \text{Var}(\mu_i + v_{it}) = \text{Var}(\mu_i) + 2\,\text{Cov}(\mu_i, v_{it}) + \text{Var}(v_{it}) = \text{Var}(\mu_i) + 0 + 1$$

We must make an a priori estimate of each $\text{Var}(\mu_i)$—and we choose 0. This lets us proceed as if $\varepsilon_{it} = v_{it}$. Then paralleling the construction for difference GMM, $\mathbf{H}$ is block diagonal with blocks

$$\text{Var}\left(\varepsilon_i^+\right) = \text{Var}\left(v_i^+\right) = \mathbf{M}_*^+\mathbf{M}_*^{+\prime} = \begin{pmatrix} \mathbf{M}_*\mathbf{M}_*' & \mathbf{M}_* \\ \mathbf{M}_*' & \mathbf{I} \end{pmatrix} \tag{26}$$

where, in the orthogonal deviations case, $\mathbf{M}_*\mathbf{M}_*' = \mathbf{I}$. This is the default value of $\mathbf{H}$ for system GMM in `xtabond2`. Current versions of Arellano and Bond's own estimation package, DPD, zero out the upper-right and lower-left quadrants of these matrices. (Doornik, Arellano, and Bond 2006). The original implementation of system GMM (Blundell and Bond 1998) used $\mathbf{H} = \mathbf{I}$. These choices are available in `xtabond2` too.

For an application, Blundell and Bond return to the employment equation, using the same dataset as in Arellano and Bond, and we follow suit. This time, the authors drop the longest (two-period) lags of employment and capital from their model, and dispense with sector-wide demand altogether. They also switch to treating wages and capital as potentially endogenous, generating GMM-style instruments for them. The `xtabond2` command line for a one-step estimate is

```
. xtabond2 n L.n L(0/1).(w k) yr*, gmmstyle(L.(n w k))
> ivstyle(yr*, equation(level)) robust small
Favoring space over speed. To switch, type or click on mata: mata set matafavor
> speed, perm.
yr1976 dropped due to collinearity
yr1984 dropped due to collinearity
Warning: Two-step estimated covariance matrix of moments is singular.
  Using a generalized inverse to calculate robust weighting matrix for Hansen
> test.
  Difference-in-Sargan statistics may be negative.

Dynamic panel-data estimation, one-step system GMM
```

(*Continued on next page*)

```
Group variable: id                              Number of obs      =       891
Time variable : year                            Number of groups   =       140
Number of instruments = 113                     Obs per group: min =         6
F(12, 139)    =   1154.36                                      avg =      6.36
Prob > F      =     0.000                                      max =         8
```

n	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
n						
L1.	.9356053	.026569	35.21	0.000	.8830737	.9881369
w						
--.	-.6309761	.1192834	-5.29	0.000	-.8668206	-.3951315
L1.	.4826203	.1383132	3.49	0.001	.2091504	.7560901
k						
--.	.4839299	.0544281	8.89	0.000	.3763159	.591544
L1.	-.4243928	.059088	-7.18	0.000	-.5412204	-.3075653
yr1977	-.0240573	.0296969	-0.81	0.419	-.0827734	.0346588
yr1978	-.0176523	.0229277	-0.77	0.443	-.0629845	.0276799
yr1979	-.0026515	.0207492	-0.13	0.899	-.0436764	.0383735
yr1980	-.0173995	.0221715	-0.78	0.434	-.0612366	.0264376
yr1981	-.0435283	.0193348	-2.25	0.026	-.0817565	-.0053
yr1982	-.0096193	.0186829	-0.51	0.607	-.0465588	.0273201
yr1983	.0038132	.0171959	0.22	0.825	-.0301861	.0378126
_cons	.5522011	.1971607	2.80	0.006	.1623793	.9420228

```
Instruments for first differences equation
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    L(1/.).(L.n L.w L.k)
Instruments for levels equation
  Standard
    _cons
    yr1976 yr1977 yr1978 yr1979 yr1980 yr1981 yr1982 yr1983 yr1984
  GMM-type (missing=0, separate instruments for each period unless collapsed)
    D.(L.n L.w L.k)
```

```
Arellano-Bond test for AR(1) in first differences: z =  -5.46  Pr > z =  0.000
Arellano-Bond test for AR(2) in first differences: z =  -0.25  Pr > z =  0.804
```

```
Sargan test of overid. restrictions: chi2(100)  = 186.90  Prob > chi2 =  0.000
  (Not robust, but not weakened by many instruments.)
Hansen test of overid. restrictions: chi2(100)  = 110.70  Prob > chi2 =  0.218
  (Robust, but can be weakened by many instruments.)

Difference-in-Hansen tests of exogeneity of instrument subsets:
  GMM instruments for levels
    Hansen test excluding group:     chi2(79)   =  84.33  Prob > chi2 =  0.320
    Difference (null H = exogenous): chi2(21)   =  26.37  Prob > chi2 =  0.193
  iv(yr1976 yr1977 yr1978 yr1979 yr1980 yr1981 yr1982 yr1983 yr1984, eq(level))
    Hansen test excluding group:     chi2(93)   = 107.79  Prob > chi2 =  0.140
    Difference (null H = exogenous): chi2(7)    =   2.91  Prob > chi2 =  0.893
```

These estimates do not match the published ones, in part because Blundell and Bond set $\mathbf{H} = \mathbf{I}$ instead of using the form in (26).[17] The new point estimate of the coefficient on lagged employment is higher than that at the end of section 3.3 though

17. One could add an `h(1)` option to the command line to mimic their choice.

not statistically different with reference to the previous standard errors. Moreover, the new coefficient estimate is within the bracketing LSDV–OLS range of 0.733–1.045, and the reported standard error is half its previous value.

3.5 Testing for autocorrelation

The Sargan/Hansen test for joint validity of the instruments is standard after GMM estimation. In addition, Arellano and Bond develop a test for a phenomenon that would render some lags invalid as instruments, namely, autocorrelation in the idiosyncratic disturbance term, v_{it}. Of course, the full disturbance, ε_{it}, is presumed autocorrelated because it contains fixed effects, and the estimators are designed to eliminate this source of trouble. But if the v_{it} are themselves serially correlated of order 1 then, for instance, $y_{i,t-2}$ is endogenous to the $v_{i,t-1}$ in the error term in differences, $\Delta\varepsilon_{it} = v_{it} - v_{i,t-1}$, making it a potentially invalid instrument after all. The researcher would need to restrict the instrument set to lags 3 and longer of y—unless the researcher found order-2 serial correlation, in which case he or she would need to start with even longer lags.

To test for autocorrelation aside from the fixed effects, the Arellano–Bond test is applied to the residuals in differences. Because Δv_{it} is mathematically related to $\Delta v_{i,t-1}$ via the shared $v_{i,t-1}$ term, negative first-order serial correlation is expected in differences and evidence of it is uninformative. Thus to check for first-order serial correlation in levels, we look for second-order correlation in differences, on the idea that this will detect correlation between the $v_{i,t-1}$ in Δv_{it} and the $v_{i,t-2}$ in $\Delta v_{i,t-2}$. In general, we check for serial correlation of order l in levels by looking for correlation of order $l+1$ in differences. Such an approach would not work for orthogonal deviations because all residuals in deviations are mathematically interrelated, depending as they do on many forward "lags". So even after estimation in deviations, the test is run on residuals in differences.

The Arellano–Bond test for autocorrelation is actually valid for any GMM regression on panel data, including OLS and 2SLS, as long as none of the regressors is "postdetermined", depending on future disturbances. (A fixed-effects or within-groups regression can violate this assumption if T is small.) Also we must assume that errors are not correlated *across* individuals.[18] The `abar` command makes the test available after `regress`, `ivregress`, `ivreg2`, `newey`, and `newey2`. So, in deriving the test, we will refer to a generic GMM estimate, $\widehat{\beta}_{\mathbf{A}}$, applied to a dataset, $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, which may have been pretransformed; the estimator yields residuals $\widehat{\mathbf{E}}$.

18. For similar reasons, the test appears appropriate for ergodic time series.

If $\mathbf{W}$ is a data matrix, let $\mathbf{W}^{-l}$ be its l lag, with zeroes for $t \leq l$. The Arellano–Bond autocorrelation test is based on the inner product $(1/N)\sum_i \widehat{\mathbf{E}}_i^{-l}\widehat{\mathbf{E}}_i$, which is zero in expectation under the null of zero order-l serial correlation. Assuming errors are sufficiently uncorrelated across individuals, a central limit theorem assures that the statistic

$$\sqrt{N}\frac{1}{N}\sum_i \widehat{\mathbf{E}}_i^{-l}\widehat{\mathbf{E}}_i = \frac{1}{\sqrt{N}}\widehat{\mathbf{E}}^{-l}\widehat{\mathbf{E}} \tag{27}$$

is asymptotically normally distributed. For the tendency toward normality to set in, only N, not T, needs be large.

To estimate the asymptotic variance of the statistic under the null, Arellano and Bond start like the Windmeijer derivation above, expressing the quantity of interest as a deviation from the theoretical value it approximates. In particular, because $\mathbf{Y} = \mathbf{X}\beta + \mathbf{E} = \mathbf{X}\widehat{\beta} + \widehat{\mathbf{E}}$, $\widehat{\mathbf{E}} = \mathbf{E} - \mathbf{X}\left(\widehat{\beta}_{\mathbf{A}} - \beta\right)$. Substituting into (27) gives

$$\begin{aligned}
\frac{1}{\sqrt{N}}\widehat{\mathbf{E}}^{-l}\widehat{\mathbf{E}} &= \frac{1}{\sqrt{N}}\left\{\mathbf{E}^{-l} - \mathbf{X}_i^{-l}\left(\widehat{\beta}_{\mathbf{A}} - \beta\right)\right\}'\left\{\mathbf{E} - \mathbf{X}\left(\widehat{\beta}_{\mathbf{A}} - \beta\right)\right\} \\
&= \frac{1}{\sqrt{N}}\mathbf{E}^{-l\prime}\mathbf{E} - \frac{\mathbf{E}^{-l\prime}\mathbf{X}}{N}\sqrt{N}\left(\widehat{\beta}_{\mathbf{A}} - \beta\right) \\
&\quad -\sqrt{N}\left(\widehat{\beta}_{\mathbf{A}} - \beta\right)'\frac{\mathbf{X}^{-l\prime}\mathbf{E}}{N} + \sqrt{N}\left(\widehat{\beta}_{\mathbf{A}} - \beta\right)'\frac{1}{\sqrt{N}}\frac{\mathbf{X}^{-l\prime}\mathbf{X}}{N}\sqrt{N}\left(\widehat{\beta}_{\mathbf{A}} - \beta\right)
\end{aligned} \tag{28}$$

The last two terms drop out as $N \to \infty$. Why? Because $\widehat{\beta}_{\mathbf{A}}$ is a $\sqrt{N}$-consistent estimate of β (Ruud 2000, 546), the $\sqrt{N}\left(\widehat{\beta}_{\mathbf{A}} - \beta\right)$ terms neither diverge nor converge to 0. Meanwhile, assuming $\mathbf{x}$ is not postdetermined, $\mathbf{X}^{-l\prime}\mathbf{E}/N$ goes to 0, which eliminates the third term. For similar reasons, assuming that $\mathbf{X}^{-l\prime}\mathbf{X}/N$ does not diverge, the fourth term goes to zero. If we then substitute (3) into the second term, the expression converges to $(1/\sqrt{N})\left\{\mathbf{E}^{-l\prime}\mathbf{E} - \mathbf{E}^{-l\prime}\mathbf{X}\left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{E}\right\}$, whose variance is consistently estimated by

$$\begin{aligned}
\frac{1}{\sqrt{N}}\Big\{&\widehat{\mathbf{E}}^{-l\prime}\widehat{\mathrm{Var}}\left(\widehat{\mathbf{E}}\,|\,\mathbf{Z}\right)\widehat{\mathbf{E}}^{-l} - 2\widehat{\mathbf{E}}^{-l\prime}\mathbf{X}\left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\widehat{\mathrm{Var}}\left(\widehat{\mathbf{E}}\,|\,\mathbf{Z}\right)\widehat{\mathbf{E}}^{-l} \\
&+\widehat{\mathbf{E}}^{-l\prime}\mathbf{X}\widehat{\mathrm{Avar}}\left(\widehat{\beta}_{\mathbf{A}}\right)\mathbf{X}'\widehat{\mathbf{E}}^{-l}\Big\}
\end{aligned}$$

(Arellano and Bond 1991). Dividing this value into (27) to normalize it yields the Arellano–Bond z test for serial correlation of order l.

For difference and system GMM, terms in this formula map as follows. $\widehat{\mathbf{E}}^{-l}$ contains lagged, differenced errors, with observations for the levels data zeroed out in system GMM because they are not the basis for the test. $\mathbf{X}$ and $\mathbf{Z}$ hold the transformed and, in system GMM, augmented dataset used in the estimation. In one-step, nonrobust estimation, $\widehat{\mathrm{Var}}\left(\widehat{\mathbf{E}}\,|\,\mathbf{Z}\right)$ is $\widehat{\sigma}^2\mathbf{H}$, where $\widehat{\sigma}$ is a consistent estimate of the standard deviation

of the errors in levels. Otherwise, $\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}$ is substituted. $\widehat{\text{Avar}}\left(\widehat{\beta}_{\mathbf{A}}\right)$ is set to the reported variance matrix—robust or not, Windmeijer-corrected or not.[19]

There are two important lessons here for the researcher. The first is another reminder of the importance of time dummies in preventing the most likely form of cross-individual correlation: contemporaneous correlation. The second lesson is that the test depends on the assumption that N is large. Large has no precise definition, but applying it to panels with $N = 20$, for instance, seems worrisome.

In their difference GMM regressions on simulated 7×100 panels with $AR(1)$, Arellano and Bond find that their test has greater power than the Sargan and Hansen tests to detect lagged instruments being made invalid through autocorrelation. However, the test does break down as the correlation falls to 0.2, where it rejects the null of no serial correlation only half the time.

4 Implementation

4.1 Syntax

The original implementation of difference GMM is the DPD package, written in the Gauss programming language (Arellano and Bond 1998). An update, DPD98, incorporates system GMM. DPD has also been implemented in the Ox language (Doornik, Arellano, and Bond 2006). In 2001, StataCorp shipped `xtabond` in Stata 7. It performed difference GMM but not system GMM nor the Windmeijer correction. In late 2003, I set out to add these features. In the end, I revamped the code and syntax and added other options. `xtabond2` was and is compatible with Stata version 7 and later. I also wrote `abar` to make the Arellano–Bond autocorrelation test available after other estimation commands. Stata 10, shipped in mid-2007, incorporated many features of `xtabond2`, via a revised `xtabond` and the new `xtdpd` and `xtdpdsys` commands. Unlike the official Stata commands, which have computationally intensive sections precompiled, the first versions of `xtabond2` were written purely in Stata's interpreted ado language, which made it slow. In late 2005, I implemented `xtabond2` afresh in the Mata language shipped with Stata 9; the Mata version runs much faster, though not as fast as the built-in commands. The two `xtabond2` implementations are bundled together, and the ado version automatically runs if Mata is not available.[20]

19. In one-step, nonrobust estimation in orthogonal deviations, the second $\widehat{\text{Var}}\left(\widehat{\mathbf{E}} \,|\, \mathbf{Z}\right)$ is actually set to $\mathbf{M}_{\perp}\mathbf{M}_{\Delta}{}'$ in difference GMM and $\mathbf{M}_{\perp}^{+}\mathbf{M}_{\Delta}^{+\prime}$ in system GMM.
20. The Mata code requires Stata 9.1 or later. Version 9.0 users will be prompted to upgrade for free.

The syntax for `xtabond2` is

`xtabond2` *depvar varlist* [*if*] [*in*] [*weight*] [`, level(`#`) twostep robust cluster(`*varname*`) noconstant small noleveleq orthogonal artests(`#`) arlevels h(`#`) nodiffsargan nomata` *ivopt* [*ivopt ...*] *gmmopt* [*gmmopt ...*]]

where *ivopt* is

`ivstyle(`*varlist* [`, equation(diff|level|both) passthru mz`])

and *gmmopt* is

`gmmstyle(`*varlist* [`, laglimits(`*a b*`) collapse equation(diff|level|both) passthru split`])

All *varlist*s can include time-series operators, such as `L.`, and wildcard expressions, such as `_I*`.

The `if` and `in` qualifiers restrict the estimation sample, but they do not restrict the sample from which lagged variables are drawn for instrument construction. *weight* also follows Stata conventions; analytical weights (`aweight`s), sampling weights (`pweight`s), and frequency weights (`fweight`s) are accepted. Frequency weights must be constant over time. (See the appendix for details.)

The `level()`, `robust`, `cluster()`, `noconstant`, and `small` options are also mostly standard. `level()` controls the size of the reported confidence intervals, the default being 95%. In one-step GMM, `xtabond2`'s `robust` option is equivalent to `cluster(id)` in most other estimation commands, where `id` is the panel identifier variable, requesting standard errors that are robust to heteroskedasticity and arbitrary patterns of autocorrelation within individuals; in two-step estimation, where the errors are already robust, `robust` triggers the Windmeijer correction. `cluster()` overrides the default use of the panel identifier (as set by `tsset`) as the basis for defining groups. It implies `robust` in the senses just described. Changing the clustering variable with this option affects one-step "robust" standard errors, all two-step results, the Hansen and difference-in-Hansen tests, and the Arellano–Bond serial correlation tests. `cluster()` is available only in the Mata version of `xtabond2`, which requires Stata 9 or later. `noconstant` excludes the constant term from $\mathbf{X}$ and $\mathbf{Z}$; however, it has no effect in difference GMM because differencing eliminates the constant anyway.[21] `small` requests small-sample corrections to the covariance matrix estimate, resulting in t-test instead of z-test statistics for the coefficients and an F test instead of a Wald χ^2 test for overall fit.

Most of the other options are straightforward. `nomata` prevents the use of the Mata implementation even when it is available, in favor of the ado program. `twostep` requests two-step FEGMM, one-step GMM being the default. `noleveleq` invokes difference instead of system GMM, which is the default. `nodiffsargan` prevents reporting of certain difference-in-Sargan/Hansen statistics (described below), which are computationally intensive because they involve reestimating the model for each test. It has effect only in the Mata implementation, because only that version performs the tests. `orthogonal`, also meaningful only for the Mata version, requests the forward orthogonal-deviations transform instead of first differencing. `artests()` sets the maximum lag distance to check for autocorrelation, the default being 2. `arlevels` requests that the Arellano–Bond autocorrelation test be run on the levels residuals instead of the differenced ones; it applies only to system GMM and makes sense only in the unconventional case where it is believed that there are no fixed effects whose own autocorrelation would mask any in the idiosyncratic errors. The `h(#)` option, which most users can safely ignore, controls the choice of $\mathbf{H}$. `h(1)` sets $\mathbf{H} = \mathbf{I}$, for both difference and system GMM. For difference GMM, `h(2)` and `h(3)` coincide, making the matrix in (22). They differ for system GMM, however, with `h(2)` imitating DPD for Ox and `h(3)` being the `xtabond2` default, according to (26) (see the end of section 3.4).

21. Here `xtabond2` differs from `xtabond`, `xtdpd`, and DPD, which by default enter the constant in difference GMM *after* transforming the data. DPD does the same for time dummies. `xtabond2` avoids this practice for several reasons. First, in Stata, it is more natural to treat time dummies, typically created with `xi`, like any other regressor, transforming them. Second, introducing the constant term after differencing is equivalent to entering t as a regressor before transformation, which may not be what users intend. By the same token, it introduces an inconsistency with system GMM: in DPD and `xtdpdsys`, when doing system GMM, the constant term enters only in the levels equation, and in the usual way; it means 1 rather than t. Thus switching between difference and system GMM changes the model. However, these problems are minor as long as a full set of time dummies is included. Because the linear span of the time dummies and the constant term together is the same as that of their first differences or orthogonal deviations, it does not matter much whether the time dummies and constant enter transformed or not.

The most important thing to understand about the `xtabond2` syntax is that unlike most Stata estimation commands, including `xtabond`, the variable list before the comma *communicates no identification information*. The first variable defines **Y** and the remaining ones define **X**. None of them say anything about **Z** even though **X** and **Z** can share columns. Designing the instrument matrix is the job of the `ivstyle()` and `gmmstyle()` options after the comma, each of which can be listed multiple times or not at all. (`noconstant` also affects **Z** in system GMM.) As a result, most regressors appear twice in a command line, once before the comma for inclusion in **X** and once after as a source of IV- or GMM-style instruments. Variables that serve only as excluded instruments appear once, in `ivstyle()` or `gmmstyle()` options after the comma.

The standard treatment for strictly exogenous regressors or IV-style excluded instruments, say, `w1` and `w2`, is `ivstyle(w1 w2)`. This generates one column per variable, with missing not replaced by 0. In particular, strictly exogenous regressors ordinarily instrument themselves, appearing in both the variable list before the comma and in an `ivstyle()` option. In difference GMM, these IV-style columns are transformed unless the user specifies `ivstyle(w1 w2, passthru)`. `ivstyle()` also generates one column per variable in system GMM, following (24). The patterns in (25) can be requested with the `equation()` suboption, as in `ivstyle(w1 w2, equation(level))` and the compound `ivstyle(w1 w2, equation(diff)) ivstyle(w1 w2, equation(level))`. The `mz` suboption instructs `xtabond2` to substitute zero for missing in the generated IV-style instruments.

Similarly, the `gmmstyle()` option includes a list of variables, then suboptions after a comma that control how the variables enter **Z**. By default, `gmmstyle()` generates the instruments appropriate for predetermined variables: lags 1 and earlier of the instrumenting variable for the transformed equation and, for system GMM, lag 0 of the instrumenting variable in differences for the levels equation. The `laglimits()` suboption overrides the defaults on lag range. For example, `gmmstyle(w, laglimits(2 .))` specifies lags 2 and longer for the transformed equation and lag 1 for the levels equation, which is the standard treatment for endogenous variables. In general, `laglimits(`a b`)` requests lags a through b of the levels as instruments for the transformed data and lag $a-1$ of the differences for the levels data. a and b can each be missing ("`.`"). a defaults to 1 and b to infinity, so that `laglimits(. .)` is equivalent to leaving the suboption out altogether. a and b can even be negative, implying forward "lags". If $a > b$, `xtabond2` swaps their values.[22] Because the `gmmstyle()` varlist allows time-series operators, there are many routes to the same specification. For example, if `w1` is predetermined and `w2` endogenous, then instead of `gmmstyle(w1) gmmstyle(w2, laglimits(2 .))`, one could simply type `gmmstyle(w1 L.w2)`. In all these instances, the suboption `collapse` is available to "collapse" the instrument sets as described in sections 3.2 and 3.4.

22. If $a \leq b < 0$, then lag $b-1$ of the differences is normally used as an instrument in the levels equations instead of that dated $a-1$, because it is more frequently in the range $[1, T]$ of valid time indexes. Or, for the same reasons, if $a \leq 0 \leq b$ or $b \leq 0 \leq a$, the contemporaneous difference is used. Tue Gørgens developed these decision rules.

`gmmstyle()` also has `equation()` and `passthru` suboptions, which work much like their `ivstyle()` counterparts. The exception is that `equation(level)`, by blocking the generation of the instruments for the transformed equation, causes `xtabond2` to generate a full GMM-style set of instruments for the levels equation because they are no longer mathematically redundant.[23] `passthru` prevents the usual differencing of instruments for the levels equation. As with `arlevels`, this produces invalid results under standard assumptions. A final suboption, `split`, is explained below.

Along with the standard estimation results, `xtabond2` reports the Sargan/Hansen test, Arellano–Bond autocorrelation tests, and various summary statistics. Sample size is not an entirely well-defined concept in system GMM, which runs in effect on two samples simultaneously. `xtabond2` reports the size of the transformed sample after difference GMM and of the untransformed sample after system GMM.

The Mata implementation carries out certain difference-in-Sargan/Hansen tests unless `nodiffsargan` is specified. In particular, it reports a difference-in-Sargan/Hansen test for each instrument group defined by an `ivstyle()` or `gmmstyle()` option, when feasible. So a clause like `gmmstyle(x y)` implicitly requests one test for this entire instrument group, while `gmmstyle(x) gmmstyle(y)` requests the same estimates but two more-targeted difference-in-Sargan/Hansen tests. In system GMM, a `split` suboption in a `gmmstyle()` option instructs `xtabond2` to subject the transformed- and levels-equation instruments within the given GMM-style group to separate difference tests. This facilitates testing of the instruments of greatest concern in system GMM, those for the levels equation based on the dependent variable. The Mata version also tests all the GMM-style instruments for the levels equation as a group.[24]

The Mata version of `xtabond2` responds to one option that is not set in the command line, namely, the Mata system parameter `matafavor`. When this is set to `speed` (which can be done by typing `mata: mata set matafavor speed, permanently` at the Stata prompt), the Mata code builds a complete internal representation of $\mathbf{Z}$.[25] If there are 1,000 observations and 100 instruments, then $\mathbf{Z}$ will contain some 200,000 elements in system GMM, each of which will take 8 bytes in Mata, for a total of roughly 1.5

23. Because an ordinary `gmmstyle(w, laglimits(`a b`))` command in system GMM requests lags a through b of `w` as instruments for the transformed equation and lag $a-1$ of Δ`w` for the levels equation, for consistency, `xtabond2`, in versions 1.2.8 and earlier, interpreted `gmmstyle(w, laglimits(`a b`) equation(level))` to request lags $a-1$ through $b-1$ of Δ`w` for the levels equation. But with version 2.0.0, the interpretation changed to lags a through b.
24. The reported differences in Sargan/Hansen will generally not match what would be obtained by manually running the estimation with and without the suspect instruments. Recall from section 2.3 that in the full, restricted regression, the moment weighting matrix is the inverse of the estimated covariance of the moments, call it $\widehat{\mathbf{S}}$, which is $\mathbf{Z}'\mathbf{H}\mathbf{Z}$ in one-step and $\mathbf{Z}'\widehat{\mathbf{\Omega}}_{\hat{\beta}_1}\mathbf{Z}$ in two-step. In the unrestricted regressions carried out for testing purposes, `xtabond2` weights using the submatrix of the restricted $\widehat{\mathbf{S}}$ corresponding to the nonsuspect instruments. This reduces the chance of a negative test statistic (Baum, Schaffer, and Stillman [2003, 18], citing Hayashi [2000]). As described in section 2.6, adding instruments weakens the Sargan/Hansen test and can actually reduce the statistic, which is what makes negative differences in Sargan/Hansen more likely if the unrestricted regression is fully reestimated.
25. Despite the `speed` setting, there is a delay the first time the Mata version of `xtabond2` runs in a Stata session, because Stata loads the function library.

megabytes. Larger panels can exceed a computer's physical memory and actually even slow Mata down because the operating system is forced to repeatedly cache parts of **Z** to the hard drive and then reload them. Setting `matafavor` to `space` causes the program to build and destroy submatrices $\mathbf{Z}_i$ for each individual "on the fly". The Mata code in this mode can be even slower than the ado version, but because the ado version also builds a full representation of **Z**, the Mata code in `space` mode still has the advantage of conserving memory.

The Mata and ado implementations should generate identical results. However, if some regressors are nearly or fully multicollinear, the two may disagree on the number and choice of regressors to drop. Because floating-point representations of numbers have finite precision, even exactly collinear variables may not quite appear that way to the computer, and algorithms for identifying them must look for "near-multicollinearity". There is no one right definition for that term, and the identification can be sensitive to the exact procedure. Where the ado program calls the built-in Stata command `_rmcoll`, the Mata program must use its own procedure, which differs in logic and tolerances.[26]

As a Stata estimation command, `xtabond2` can be followed by `predict`:

`predict` [*type*] *newvarname* [*if*] [*in*] [`,` *statistic* `difference`]

where *statistic* is `xb` or `residuals`. The optional *type* clause controls the data type of the variable generated. Requesting the `xb` statistic, the default, essentially gives $\mathbf{X}\widehat{\beta}$, where $\widehat{\beta}$ is the parameter vector from the estimation. However, difference GMM never estimates a coefficient on the constant term, so `predict` can predict the dependent variable only up to a constant. To compensate, after difference GMM `predict` adds a constant to the series chosen to give it the same average as **Y**. Putting `residuals` in the command line requests $\mathbf{Y} - \mathbf{X}\widehat{\beta}$, where the $\mathbf{X}\widehat{\beta}$ again will be adjusted. The `difference` option requests predictions and residuals in differences.

The syntax for the postestimation command `abar` is

`abar` [*if*] [*in*] [`, lags(`#`)`]

The `lags()` option works like `xtabond2`'s `artests()` option except that it defaults to 1. `abar` can run after `regress`, `ivregress`, `ivreg2`, `newey`, and `newey2`. It tests for autocorrelation in the estimation errors, undifferenced.

4.2 More examples

A simple autoregressive model with no controls except time dummies would be fit by

```
. xi: xtabond2 y L.y i.t, gmmstyle(L.y) ivstyle(i.t) robust noleveleq
```

26. The Mata version will not perfectly handle strange and unusual expressions like `gmmstyle(L.x, laglimits(-1 -1))`. This is the same as `gmmstyle(x, laglimits(0 0))` in principle. But the Mata code will interpret it by lagging `x`, thus losing the observation of `x` for $t = T$, and then unlagging the remaining information. The ado version does not lose data in this way.

where `t` is the time variable. This would run one-step difference GMM with robust errors. If `w1` is strictly exogenous, `w2` is predetermined but not strictly exogenous, and `w3` is endogenous, then

```
. xi: xtabond2 y L.y w1 w2 w3 i.t, gmmstyle(L.y w2 L.w3) ivstyle(i.t w1)
> twostep robust small orthogonal
```

would fit the model with the standard choices of instruments—here with two-step system GMM, Windmeijer-corrected standard errors, small-sample adjustments, and orthogonal deviations.

If the user runs system GMM without declaring instruments that are nonzero for the transformed equation, then the estimation is effectively run on levels only. Moreover, though it is designed for dynamic models, `xtabond2` does not require the lagged dependent variable to appear on the right-hand side. As a result, the command can perform OLS and 2SLS. Following are pairs of equivalents, all of which can be run on the Arellano–Bond dataset:

```
. regress n w k
. abar
. xtabond2 n w k, ivstyle(w k, equation(level)) small arlevels artests(1)

. ivreg2 n cap (w = k ys), cluster(id)
. abar, lags(2)
. xtabond2 n w cap, ivstyle(cap k ys, equation(level)) small robust arlevels

. ivreg2 n cap (w = k ys), cluster(id) gmm
. abar
. xtabond2 n w cap, ivstyle(cap k ys, equation(level)) twostep artests(1) arlevels
```

The only value in such tricks is that they make the Windmeijer correction available for linear GMM regressions more generally.

`xtabond2` can replicate results from comparable packages. Here is a matching triplet:

```
. xtabond n, lags(1) pre(w, lagstruct(1,.)) pre(k, endog) robust
. xtdpd n L.n w L.w k, dgmmiv(w k n) vce(robust)
. xtabond2 n L.n w L.w k, gmmstyle(L.(w n k), eq(diff)) robust
```

To exactly match difference GMM results from DPD for Gauss and Ox, one must also create variables that become the constant and time dummies *after* transformation, to mimic the way DPD enters these variables directly into the difference equation. This example exactly imitates the regression for column a1 in table 4 of Arellano and Bond (1991):

```
.forvalues y = 1979/1984 { /* Make variables whose differences are time dummies */
2.    gen yr`y´c = year>=`y´
3. }
. gen cons = year
. xtabond2 n L(0/1).(L.n w) L(0/2).(k ys) yr198?c cons, gmmstyle(L.n)
> ivstyle(L(0/1).w L(0/2).(k ys) yr198?c cons) noleveleq noconstant small
> robust
```

For system GMM, these gymnastics are unnecessary because DPD enters the constant and time dummies directly into the levels equation, not the difference one. These two

commands exactly reproduce a version of Blundell and Bond's (1998) regression 4, table 4, included in a demonstration file shipped with DPD for Ox:[27]

```
. xtdpd n L.n L(0/1).(w k) yr1978-yr1984, dgmm(w k n) lgmm(w k n)
> liv(yr1978-yr1984) vce(robust) two hascons
. xtabond2 n L.n L(0/1).(w k) yr1978-yr1984, gmmstyle(L.(w k n))
> ivstyle(yr1978-yr1984, equation(level)) h(2) robust twostep small
```

More replications from the regressions in the Arellano and Bond (1998) and Blundell and Bond (1998) articles are in two ancillary files that come with `xtabond2`: `abest.do` and `bbest.do`. In addition, `greene.do` reproduces an example in Greene (2003, 554).[28]

5 Conclusion

By way of conclusion, I offer a few pointers on the use of difference and system GMM, however implemented. Most of these are discussed above.

- *Apply the estimators to "small T, large N" panels.* If T is large, dynamic panel bias becomes insignificant, and a more straightforward fixed-effects estimator works. Meanwhile, the number of instruments in difference and system GMM tends to explode with T. If N is small, the cluster–robust standard errors and the Arellano–Bond autocorrelation test may be unreliable.

- *Include time dummies.* The autocorrelation test and the robust estimates of the coefficient standard errors assume no correlation across individuals in the idiosyncratic disturbances. Time dummies make this assumption more likely to hold.

- *Use orthogonal deviations in panels with gaps.* This maximizes sample size.

- *Ordinarily, put every regressor into the instrument matrix, $\mathbf{Z}$, in some form.* If a regressor, w, is strictly exogenous, standard treatment is to insert it as one column (in `xtabond2`, with `ivstyle(w)`). If `w` is predetermined to not be strictly exogenous, standard treatment is to use lags 1 and longer, GMM-style (`gmmstyle(w)`). And if `w` is endogenous, standard treatment is lags 2 and longer (`gmmstyle(L.w)`).

- *Before using system GMM, ponder the required assumptions.* The validity of the additional instruments in system GMM depends on the assumption that changes in the instrumenting variables are uncorrelated with the fixed effects. In particular, they require that throughout the study period, individuals sampled are not too far from steady states, in the sense that deviations from long-run means are not systematically related to fixed effects.

- *Mind and report the instrument count.* As discussed in section 2.6 and Roodman (2009), instrument proliferation can overfit endogenous variables and fail to expunge their endogenous components. Ironically, it also weakens the power of

27. In the command file `bbest.ox`.
28. To download them into your current directory, type `net get xtabond2` in Stata.

the Hansen test to detect this very problem and to detect invalidity of the system GMM instruments, whose validity should not be taken for granted. Because the risk is high with these estimators, researchers should report the number of instruments and reviewers should question regressions where it is not reported. A telltale sign is a perfect Hansen statistic of 1.000. Researchers should also test for robustness to severely reducing the instrument count. Options include limiting the lags used in GMM-style instruments and, in `xtabond2`, collapsing instruments. Also, because of the risks, do not take comfort in a Hansen test p-value below 0.1. View higher values, such as 0.25, as potential signs of trouble.

- *Report all specification choices.* Using these estimators involves many choices, and researchers should report the ones they make—difference or system GMM; first differences or orthogonal deviations; one- or two-step estimation; nonrobust, cluster–robust, or Windmeijer-corrected cluster–robust errors; and the choice of instrumenting variables and lags used.

6 Acknowledgments

I thank Manuel Arellano, Christopher Baum, Michael Clemens, Decio Coviello, Mead Over, Mark Schaffer, and one anonymous reviewer for comments. I also thank all the users whose feedback has led to steady improvement in `xtabond2`.

7 References

Anderson, T. G., and B. E. Sørenson. 1996. GMM estimation of a stochastic volatility model: A Monte Carlo study. *Journal of Business & Economic Statistics* 328–352.

Anderson, T. W., and C. Hsiao. 1982. Formulation and estimation of dynamic models using panel data. *Journal of Econometrics* 18: 47–82.

Arellano, M., and S. Bond. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58: 277–297.

———. 1998. Dynamic panel data estimation using DPD98 for Gauss: A guide for users. Mimeo. Available at ftp://ftp.cemfi.es/pdf/papers/ma/dpd98.pdf.

Arellano, M., and O. Bover. 1995. Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics* 68: 29–51.

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

Beck, T., and R. Levine. 2004. Stock markets, banks, and growth: Panel evidence. *Journal of Banking and Finance* 28: 423–442.

Blundell, R., and S. Bond. 1998. Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87: 115–143.

Bond, S. 2002. Dynamic panel data models: A guide to micro data methods and practice. Working Paper CWP09/02, Cemmap, Institute for Fiscal Studies. Available at http://cemmap.ifs.org.uk/wps/cwp0209.pdf.

Bowsher, C. G. 2002. On testing overidentifying restrictions in dynamic panel data models. *Economics Letters* 77: 211–220.

Calderón, C. A., A. Chong, and N. V. Loayza. 2002. Determinants of current account deficits in developing countries. *Contributions to Macroeconomics* 2: Article 2.

Doornik, J. A., M. Arellano, and S. Bond. 2006. Panel data estimation using DPD for Ox. Available at http://www.doornik.com/download/dpd.pdf.

Greene, W. H. 2003. *Econometric Analysis*. 5th ed. Upper Saddle River, NJ: Prentice Hall.

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

Hayashi, F. 2000. *Econometrics*. Princeton, NJ: Princeton University Press.

Holtz-Eakin, D., W. Newey, and H. S. Rosen. 1988. Estimating vector autoregressions with panel data. *Econometrica* 56: 1371–1395.

Judson, R. A., and A. L. Owen. 1999. Estimating dynamic panel data models: A guide for macroeconomists. *Economics Letters* 65: 9–15.

Kiviet, J. F. 1995. On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. *Journal of Econometrics* 68: 53–78.

Moran, T. H., E. M. Graham, and M. Blomström. 2005. *Does Foreign Direct Investment Promote Development?* Washington, DC: Institute for International Economics.

Nickell, S. J. 1981. Biases in dynamic models with fixed effects. *Econometrica* 49: 1417–1426.

Roodman, D. M. 2009. A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics* 71: 135–158.

Ruud, P. A. 2000. *An Introduction to Classical Econometric Theory*. Oxford: Oxford University Press.

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

Windmeijer, F. 2005. A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics* 126: 25–51.

About the author

David Roodman is a research fellow at the Center for Global Development in Washington, DC.

Appendix. Incorporating observation weights

This appendix shows how weights enter the equations for GMM, the Windmeijer correction, and the Arellano–Bond autocorrelation test. Along the way, it fills a gap in section 2.4 in the derivation of the Windmeijer correction.

Stata supports several kinds of weights. Each has a different conceptual basis, but the implementations work out to be almost identical. In contrast to the matrix $\mathbf{A}$ used in the main text to weight *moments*, the weights discussed here apply at the level of *observations*. They are assumed to be exogenous.

A.1 Analytical weights

The premise of "analytical weights" (a term coined by Stata), or `aweights`, is that each observation is an average of some varying number of underlying data points. For example, the observations might be average reading scores in classes of different sizes. If the errors in the underlying data are homoskedastic, then those of the observations will not be but rather will have variance inversely proportional to the number of data points averaged. Weighting by that number is a classic way to restore homoskedasticity, thus efficiency for the coefficient estimates and consistency for the standard-error estimates.

Introducing analytical weights starts with two changes to the exposition of linear GMM in section 2. Let w be the weighting variable, assumed to be exogenous, and $\mathbf{W}$ be a diagonal $N \times N$ matrix holding the weights, normalized so that they sum to N. (Here, as in section 2, N is number of observations, not individuals.) First, the GMM criterion function in (1) becomes

$$\begin{aligned} \| E_N(\mathbf{z}\varepsilon) \|_{\mathbf{A}} &= \| \frac{1}{N}\mathbf{Z}'\mathbf{W}\widehat{\mathbf{E}} \|_{\mathbf{A}} \equiv N \left(\frac{1}{N}\mathbf{Z}'\mathbf{W}\widehat{\mathbf{E}}\right)' \mathbf{A} \left(\frac{1}{N}\mathbf{Z}'\mathbf{W}\widehat{\mathbf{E}}\right) \\ &= \frac{1}{N}\widehat{\mathbf{E}}'\mathbf{W}\mathbf{Z}\mathbf{A}\mathbf{Z}\mathbf{W}'\widehat{\mathbf{E}} \end{aligned} \tag{1'}$$

Following the derivation in the main text, this implies the weighted GMM estimator,

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{W}\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{W}\mathbf{Y} \tag{2'}$$

A proof like that in section 2.2 shows that the efficient choice of $\mathbf{A}$ is

$$\mathbf{A}_{\text{EGMM}} = \text{Var}(\mathbf{z}\mathbf{w}\varepsilon) = \text{Avar}\left(\frac{1}{N}\mathbf{Z}'\mathbf{W}\mathbf{E}\right)^{-1} \tag{4'}$$

so that EGMM is given by

$$\widehat{\beta}_{\text{EGMM}} = \left\{\mathbf{X}'\mathbf{W}\mathbf{Z}\,\text{Var}(\mathbf{z}\mathbf{w}\varepsilon)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{X}\right\}^{-1}\mathbf{X}'\mathbf{W}\mathbf{Z}\,\text{Var}(\mathbf{z}\mathbf{w}\varepsilon)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{Y} \tag{5'}$$

and FEGMM in general by

$$\widehat{\beta}_{\text{FEGMM}} = \left\{ \mathbf{X}'\mathbf{W}\mathbf{Z} \left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}\mathbf{W}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{W}\mathbf{X} \right\}^{-1} \mathbf{X}'\mathbf{W}\mathbf{Z} \left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}\mathbf{W}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{W}\mathbf{Y} \tag{11'}$$

For analytical weights, we then introduce a second change to the math. For the first step in FEGMM, we incorporate the heteroskedasticity assumption by replacing $\mathbf{H}$, the arbitrary initial approximation of $\mathbf{\Omega}$ based on the assumption of homoskedasticity, with $\mathbf{W}^{-(1/2)}\mathbf{H}\mathbf{W}^{(1/2)}$. As a result, the recipe for FEGMM becomes

$$\begin{aligned}
\widehat{\beta}_1 &= \left\{ \mathbf{X}'\mathbf{W}\mathbf{Z} \left(\mathbf{Z}'\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{W}\mathbf{X} \right\}^{-1} \mathbf{X}'\mathbf{W}\mathbf{Z} \left(\mathbf{Z}'\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{W}\mathbf{Y} \\
\widehat{\beta}_2 &= \widehat{\beta}_{\text{FEGMM}} \\
&= \left\{ \mathbf{X}'\mathbf{W}\mathbf{Z} \left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{W}\mathbf{X} \right\}^{-1} \mathbf{X}'\mathbf{W}\mathbf{Z} \left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{W}\mathbf{Y}
\end{aligned} \tag{13'}$$

For 2SLS, where $\mathbf{H} = \mathbf{I}$, $\widehat{\beta}_1$ is efficient, and simplifies to

$$\widehat{\beta}_{\text{2SLS}} = \left\{ \mathbf{X}'\mathbf{W}\mathbf{Z} \left(\mathbf{Z}'\mathbf{W}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{W}\mathbf{X} \right\}^{-1} \mathbf{X}'\mathbf{W}\mathbf{Z} \left(\mathbf{Z}'\mathbf{W}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{W}\mathbf{Y}$$

The classical variance estimate for the one- or two-step estimator is

$$\widehat{\text{Var}} \left(\widehat{\beta} \,|\, \mathbf{Z} \right) = \left\{ \mathbf{X}'\mathbf{W}\mathbf{Z} \left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}\mathbf{W}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{W}\mathbf{X} \right\}^{-1} \tag{29}$$

where $\widehat{\Omega} = \mathbf{W}^{-(1/2)}\mathbf{H}\mathbf{W}^{-(1/2)}$ or $\widehat{\Omega}_{\widehat{\beta}_1}$. And the robust one-step estimator is given by a typical sandwich formula:

$$\begin{aligned}
\widehat{\text{Var}}_{\text{r}} \left(\widehat{\beta}_1 \right) = \widehat{\text{Var}} \left(\widehat{\beta} \right) \cdot \mathbf{X}'\mathbf{W}\mathbf{Z} \left(\mathbf{Z}'\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z} \\
\times \left(\mathbf{Z}'\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{W}\mathbf{X} \cdot \widehat{\text{Var}} \left(\widehat{\beta} \right)
\end{aligned} \tag{15'}$$

Paralleling the main text, the Windmeijer correction is derived as follows:

$$\begin{aligned}
g \left(\mathbf{Y}, \widehat{\mathbf{\Omega}} \right) \equiv \left\{ \mathbf{X}'\mathbf{W}\mathbf{Z} \left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}\mathbf{W}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{W}\mathbf{X} \right\}^{-1} \\
\times \mathbf{X}'\mathbf{W}\mathbf{Z} \left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}\mathbf{W}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{W}\mathbf{E}
\end{aligned} \tag{16'}$$

$$\approx g \left(\mathbf{Y}, \widehat{\mathbf{\Omega}}_{\beta} \right) + \mathbf{D}g \left(\mathbf{Y}, \mathbf{W}^{-\frac{1}{2}}\mathbf{H}\mathbf{W}^{-\frac{1}{2}} \right) \tag{17'}$$

where $\mathbf{D} = \partial g\left(\mathbf{Y}, \widehat{\mathbf{\Omega}}_{\widehat{\beta}}\right) / \partial \widehat{\beta} \,|_{\,\widehat{\beta}=\beta}$. Repeatedly applying the identity $\partial(\mathbf{A}^{-1})/\partial b = -\mathbf{A}^{-1} \cdot \partial \mathbf{A}/\partial b \cdot \mathbf{A}^{-1}$, $\mathbf{D}$ is the matrix whose pth column is

$$-\left\{\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\beta}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{X}\right\}^{-1}\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\beta}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\frac{\partial\widehat{\mathbf{\Omega}}_{\widehat{\beta}}}{\partial\widehat{\beta}_p}\,|_{\,\widehat{\beta}=\beta}$$
$$\times\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\beta}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{E}$$

$$+\left\{\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\beta}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{X}\right\}^{-1}\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\beta}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\frac{\partial\widehat{\mathbf{\Omega}}_{\widehat{\beta}}}{\partial\widehat{\beta}_p}\,|_{\,\widehat{\beta}=\beta}$$
$$\times\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\beta}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{X}\cdot g\left(\mathbf{Y}, \widehat{\mathbf{\Omega}}\right)$$

For feasibility, Windmeijer (2005) substitutes $\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}$ for $\widehat{\mathbf{\Omega}}_{\beta}$ in this, $\widehat{\beta}_1$ for β, and $\widehat{\mathbf{E}}_2$ for $\mathbf{E}$. As a result, $g\left(\mathbf{Y}, \widehat{\mathbf{\Omega}}\right)$ becomes

$$\left\{\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{X}\right\}^{-1}\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\widehat{\mathbf{E}}_2$$

which is the projection of the two-step residuals by the two-step estimator and is exactly zero. So the second term falls out and the feasible approximation $\widehat{\mathbf{D}}$ is the matrix whose pth column is

$$-\left\{\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{X}\right\}^{-1}\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\frac{\partial\widehat{\mathbf{\Omega}}_{\widehat{\beta}}}{\partial\widehat{\beta}_p}\,|_{\,\widehat{\beta}=\widehat{\beta}_1}$$
$$\times\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\widehat{\mathbf{E}}_2$$

With this approximation in hand, we turn to estimating the asymptotic variance of (17′). For compactness, we note, using (29), that

$$g\left(\mathbf{Y}, \widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\right) = \widehat{\text{Avar}}\left(\widehat{\beta}_2\right)\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\mathbf{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{E}$$
$$g\left(\mathbf{Y}, \mathbf{W}^{-\frac{1}{2}}\mathbf{H}\mathbf{W}^{-\frac{1}{2}}\right) = \widehat{\text{Avar}}\left(\widehat{\beta}_1\right)\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{E}$$

(*Continued on next page*)

So

$$\begin{aligned}
&\operatorname{Avar}\left\{g\left(\mathbf{Y},\widehat{\boldsymbol{\Omega}}_{\widehat{\beta}_1}\right)+\widehat{\mathbf{D}}g\left(\mathbf{Y},\mathbf{W}^{-\frac{1}{2}}\mathbf{H}\mathbf{W}^{-\frac{1}{2}}\right)\right\}\\
&\quad=\operatorname{Avar}\left\{\widehat{\operatorname{Avar}}\left(\widehat{\beta}_2\right)\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\boldsymbol{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{E}+\widehat{\mathbf{D}}\widehat{\operatorname{Avar}}\left(\widehat{\beta}_1\right)\mathbf{X}'\mathbf{W}\mathbf{Z}\right.\\
&\qquad\left.\times\left(\mathbf{Z}'\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{E}\right\}\\
&\quad=\widehat{\operatorname{Avar}}\left(\widehat{\beta}_2\right)\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\boldsymbol{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z}\right)^{-1}N^2\operatorname{Var}\left(\mathbf{z}\mathbf{w}\varepsilon\right)\left(\mathbf{Z}'\mathbf{W}\widehat{\boldsymbol{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z}\right)^{-1}\\
&\qquad\times\mathbf{Z}'\mathbf{W}\mathbf{X}\widehat{\operatorname{Avar}}\left(\widehat{\beta}_2\right)+\widehat{\operatorname{Avar}}\left(\widehat{\beta}_2\right)\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\boldsymbol{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z}\right)^{-1}N^2\\
&\qquad\times\operatorname{Var}\left(\mathbf{z}\mathbf{w}\varepsilon\right)\left(\mathbf{Z}'\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{X}\widehat{\operatorname{Avar}}\left(\widehat{\beta}_1\right)\widehat{\mathbf{D}}\\
&\qquad+\widehat{\mathbf{D}}\widehat{\operatorname{Avar}}\left(\widehat{\beta}_1\right)\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Z}\right)^{-1}N^2\operatorname{Var}\left(\mathbf{z}\mathbf{w}\varepsilon\right)\left(\mathbf{Z}'\mathbf{W}\widehat{\boldsymbol{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z}\right)^{-1}\\
&\qquad\times\mathbf{Z}'\mathbf{W}\mathbf{X}\widehat{\operatorname{Avar}}\left(\widehat{\beta}_1\right)+\widehat{\mathbf{D}}\widehat{\operatorname{Avar}}\left(\widehat{\beta}_1\right)\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Z}\right)^{-1}N^2\\
&\qquad\times\operatorname{Var}\left(\mathbf{z}\mathbf{w}\varepsilon\right)\left(\mathbf{Z}'\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{X}\widehat{\operatorname{Avar}}\left(\widehat{\beta}_1\right)\widehat{\mathbf{D}}
\end{aligned}$$

Substituting once more feasibility—replacing $N^2\operatorname{Var}\left(\mathbf{z}\mathbf{w}\varepsilon\right)$ with $\mathbf{Z}'\mathbf{W}\widehat{\boldsymbol{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z}$—then simplifying and substituting with (29) and (15′) leads to the Windmeijer correction, as in (18):

$$
\begin{aligned}
&\widehat{\text{Avar}}\left\{g\left(\mathbf{Y},\widehat{\boldsymbol{\Omega}}_{\widehat{\beta}_1}\right)+\widehat{\mathbf{D}}g\left(\mathbf{Y},\mathbf{W}^{-\frac{1}{2}}\mathbf{H}\mathbf{W}^{-\frac{1}{2}}\right)\right\}\\
&\quad=\widehat{\text{Avar}}\left(\widehat{\beta}_2\right)\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}\widehat{\boldsymbol{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{X}\widehat{\text{Avar}}\left(\widehat{\beta}_2\right)+\widehat{\text{Avar}}\left(\widehat{\beta}_2\right)\mathbf{X}'\mathbf{W}\mathbf{Z}\\
&\qquad\times\left(\mathbf{Z}'\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{X}\widehat{\text{Avar}}\left(\widehat{\beta}_1\right)\widehat{\mathbf{D}}\\
&\qquad+\widehat{\mathbf{D}}\widehat{\text{Avar}}\left(\widehat{\beta}_1\right)\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\mathbf{X}\widehat{\text{Avar}}\left(\widehat{\beta}_1\right)\\
&\qquad+\widehat{\mathbf{D}}\widehat{\text{Avar}}\left(\widehat{\beta}_1\right)\mathbf{X}'\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}\widehat{\boldsymbol{\Omega}}_{\widehat{\beta}_1}\mathbf{W}\mathbf{Z}\left(\mathbf{Z}'\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Z}\right)^{-1}\\
&\qquad\times\mathbf{Z}'\mathbf{W}\mathbf{X}\widehat{\text{Avar}}\left(\widehat{\beta}_1\right)\widehat{\mathbf{D}}\\
&\quad=\widehat{\text{Avar}}\left(\widehat{\beta}_2\right)\widehat{\text{Avar}}\left(\widehat{\beta}_2\right)^{-1}\widehat{\text{Avar}}\left(\widehat{\beta}_2\right)+\widehat{\text{Avar}}\left(\widehat{\beta}_2\right)\widehat{\text{Avar}}\left(\widehat{\beta}_1\right)^{-1}\widehat{\text{Avar}}\left(\widehat{\beta}_1\right)\widehat{\mathbf{D}}\\
&\qquad+\widehat{\mathbf{D}}\widehat{\text{Avar}}\left(\widehat{\beta}_1\right)\widehat{\text{Avar}}\left(\widehat{\beta}_1\right)^{-1}\widehat{\text{Avar}}\left(\widehat{\beta}_1\right)+\widehat{\mathbf{D}}\widehat{\text{Avar}}_{\text{r}}\left(\widehat{\beta}_2\right)\widehat{\mathbf{D}}\\
&\quad=\widehat{\text{Avar}}\left(\widehat{\beta}_2\right)+\widehat{\text{Avar}}\left(\widehat{\beta}_2\right)\widehat{\mathbf{D}}+\widehat{\mathbf{D}}\widehat{\text{Avar}}\left(\widehat{\beta}_2\right)+\widehat{\mathbf{D}}\widehat{\text{Avar}}_{\text{r}}\left(\widehat{\beta}_2\right)\widehat{\mathbf{D}}
\end{aligned}
$$

Finally, we derive a weighted version of the Arellano–Bond autocorrelation test. As in section 3.5, N is the number of individuals in a panel. Paralleling (28), the basic test statistic is

$$
\begin{aligned}
\frac{1}{\sqrt{N}}\widehat{\mathbf{E}}^{-l}\mathbf{W}\widehat{\mathbf{E}} &= \frac{1}{\sqrt{N}}\left\{\mathbf{E}^{-l}-\mathbf{X}_i^{-l}\left(\widehat{\beta}_{\mathbf{A}}-\beta\right)\right\}'\mathbf{W}\left\{\mathbf{E}-\mathbf{X}\left(\widehat{\beta}_{\mathbf{A}}-\beta\right)\right\}\\
&= \frac{1}{\sqrt{N}}\mathbf{E}^{-l\prime}\mathbf{W}\mathbf{E}-\frac{\mathbf{E}^{-l\prime}\mathbf{W}\mathbf{X}}{N}\sqrt{N}\left(\widehat{\beta}_{\mathbf{A}}-\beta\right)-\sqrt{N}\left(\widehat{\beta}_{\mathbf{A}}-\beta\right)'\\
&\qquad\times\frac{\mathbf{X}^{-l\prime}\mathbf{W}\mathbf{E}}{N}+\sqrt{N}\left(\widehat{\beta}_{\mathbf{A}}-\beta\right)'\frac{1}{\sqrt{N}}\frac{\mathbf{X}^{-l\prime}\mathbf{W}\mathbf{X}}{N}\sqrt{N}\left(\widehat{\beta}_{\mathbf{A}}-\beta\right)
\end{aligned}
$$

Assuming that the weights (before being normalized to sum to NT) do not diverge, the last two terms still drop out asymptotically, and the variance of the expression is estimated by

$$
\begin{aligned}
&\frac{1}{\sqrt{N}}\left\{\widehat{\mathbf{E}}^{-l\prime}\mathbf{W}\widehat{\text{Var}}\left(\widehat{\mathbf{E}}\,|\,\mathbf{Z}\right)\mathbf{W}\widehat{\mathbf{E}}^{-l}-2\widehat{\mathbf{E}}^{-l\prime}\mathbf{W}\mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{Z}\mathbf{A}\mathbf{Z}\mathbf{W}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\mathbf{Z}\mathbf{A}\mathbf{Z}\mathbf{W}'\right.\\
&\qquad\left.\times\widehat{\text{Var}}\left(\widehat{\mathbf{E}}\,|\,\mathbf{Z}\right)\mathbf{W}\widehat{\mathbf{E}}^{-l}+\widehat{\mathbf{E}}^{-l\prime}\mathbf{W}\mathbf{X}\widehat{\text{Avar}}\left(\widehat{\beta}_{\mathbf{A}}\right)\mathbf{X}'\mathbf{W}\widehat{\mathbf{E}}^{-l}\right\}
\end{aligned}
$$

A.2 Sampling weights

Sampling weights are used to adjust for under-sampling and over-sampling in surveys. Giving higher weight to observations that correspond to larger fractions of the population can increase efficiency. Unlike with analytical weights, there is, in general, no

assumption that the survey design introduces heteroskedasticity. In principle, then, just one premise changes in moving from analytical to sampling weights: where the assumption before in formulating the one-step estimator was that $\mathbf{\Omega} = \mathbf{W}^{-1}$, now we assume that $\mathbf{\Omega}$ is scalar. Substituting $\widehat{\mathbf{\Omega}} = \mathbf{I}$ into (11′) would redefine the one-step estimator as

$$\widehat{\beta}_1 = \left\{ \mathbf{X}'\mathbf{W}\mathbf{Z} \left(\mathbf{Z}'\mathbf{W}\mathbf{H}\mathbf{W}\mathbf{Z}\right)^{-1} \mathbf{Z}'\mathbf{W}\mathbf{X} \right\}^{-1} \mathbf{X}'\mathbf{W}\mathbf{Z} \left(\mathbf{Z}'\mathbf{W}\mathbf{H}\mathbf{W}\mathbf{Z}\right)^{-1} \mathbf{Z}'\mathbf{W}\mathbf{Y} \qquad (13'')$$

However, the Stata convention is *not* to make this change, but rather to employ exactly the same formulas as for analytical weights. Using this arguably less accurate model of the errors in the first stage does not affect the consistency of coefficient estimates—even without weights, coefficient estimates would be consistent—but it can reduce efficiency in the first stage, and it makes the classical standard-error estimates inconsistent. This may be one reason why in Stata `pweights` always trigger the `robust` option. (For `xtabond2`, "robust" means clustered errors.)

In the context of two-stage FEGMM, carrying over the formulas for analytical weighting to sample weighting in the first stage poses little problem. Recall that the first-stage proxy for $\mathbf{\Omega}$ is not assumed to be accurate, and it inevitably contains some arbitrariness.

A.3 Frequency weights

The conceptual model behind frequency weights, or `fweights`, is rather different and straightforward. A frequency weight is used to collapse duplicate observations into one, more-weighted observation to economize on memory and processing power. The estimation formulas for frequency-weighted data must, therefore, have the property that they produce the same answer when run on an expanded, unweighted version of the data. The formulas for analytical weights, in fact, do behave this way, with only minor modifications. In the construction of $\mathbf{W}$, because the sum of the frequency weights *is* the true sample size N, the weights need not be normalized to sum to the number of rows in the dataset.