How to Do xtabond2:
An Introduction to "Difference" and "System" GMM in Stata
By David Roodman

Abstract

The Arellano-Bond (1991) and Arellano-Bover (1995)/Blundell-Bond (1998) linear generalized method of moments (GMM) estimators are increasingly popular. Both are general estimators designed for situations with "small T, large N" panels, meaning few time periods and many individuals; with independent variables that are not strictly exogenous, meaning correlated with past and possibly current realizations of the error; with fixed effects; and with heteroskedasticity and autocorrelation within individuals. This pedagogic paper first introduces linear GMM. Then it shows how limited time span and the potential for fixed effects and endogenous regressors drive the design of the estimators of interest, offering Stata-based examples along the way. Next it shows 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.

www.cgdev.org

How to Do xtabond2: An Introduction to "Difference" and "System" GMM in Stata[1]

David Roodman

December 2006, revised January 2007

Foreword

The Center for Global Development builds its policy recommendations on both theoretical and empirical analysis. The empirical analysis often draws on historical, non-experimental data, drawing inferences that rely partly on good judgment and partly on the considered application of the most sophisticated and appropriate econometric techniques available. This paper provides an introduction to econometric techniques that are specifically designed to extract causal lessons from data on a large number of individuals (whether countries, firms, or people) each of which is observed only a few times, such as annually over five or ten years. These techniques were developed in the 1990s by authors such as Manuel Arellano, Richard Blundell, and Olympia Bover, and have been widely applied to estimate every thing from the impact of foreign aid to the importance of financial sector development to the effects of AIDS deaths on households.

The present paper contributes to this literature pedagogically, by providing an original synthesis and exposition of the literature on these "dynamic panel estimators," and practically, by presenting the first implementation of some of these techniques in Stata, a statistical software package widely used in the research community. Stata is designed to encourage users to develop new commands for it, which other users can then use or even modify. David Roodman's **xtabond2**, introduced here, is now one of the most frequently downloaded user-written Stata commands in the world. Stata's partially open-source architecture has encouraged the growth of a vibrant world-wide community of researchers, which benefits not only from improvements made to Stata by the parent corporation, but also from the voluntary contributions of other users. Stata is arguably one of the best examples of a combination of private for-profit incentives and voluntary open-source incentives in the joint creation of a global public good. The Center for Global Development is pleased to contribute this paper and two commands, called **xtabond2** and **abar**, to the research community.

Nancy Birdsall
President
Center for Global Development

Abstract

The "difference" and "system" generalized method of moments (GMM) estimators, developed by Holtz-Eakin, Newey, and Rosen (1988), Arellano and Bond (1991), Arellano and Bover (1995), and Blundell and Bond (1998), are increasingly popular. Both are general estimators designed for situations with "small T, large N" panels, meaning few time periods and many individuals; with independent variables that are not strictly exogenous, meaning correlated with past and possibly current realizations of the error; with fixed effects; and with heteroskedasticity and autocorrelation within individuals. This pedagogic paper first introduces linear GMM. Then it shows how limited time span and the potential for fixed effects and endogenous regressors drive the design of the estimators of interest, offering Stata-based examples along the way. Next it shows 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 paper closes with some tips for proper use.

1 Introduction

The Arellano-Bond (1991) and Arellano-Bover (1995)/Blundell-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) a single left-hand-side variable that is dynamic, depending on its own past realizations; 4) independent variables that are not strictly exogneous, meaning 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 (Hansen 1982), and so is called "difference GMM."footnoteAs we will discuss, the forward orthogonal deviations transform, proposed by Arellano and Bover (1995), is sometimes performed instead of differencing. The Arellano-Bover/Blundell-Bond estimator augments Arellano-Bond by making an additional assumption, that first differences of instrumenting 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 as well as the transformed one—and is known as "system GMM." The program `xtabond2` implements these estimators. It has some important advantages over Stata's built-in `xtabond`. It implements system GMM. It can make 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 offers forward orthogonal deviations, an alternative to differencing that preserves sample size in panels with gaps. And it allows finer control over the instrument matrix.

Interestingly, though the Arellano and Bond paper 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`, like `xtabond`, automatically reports this test. But since ordinary least squares (OLS) and two-stage least squares (2SLS) are special cases of linear GMM, the Arellano-Bond test has wider applicability. The post-estimation command `abar`, also introduced in this paper, makes the test available after `regress`, `ivreg`, `ivreg2`, `newey`, and `newey2`.

One disadvantage of difference and system GMM is that they are complicated and 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

them. This paper aims to prevent that. Its approach is therefore pedagogic. Section 2 introduces linear GMM. Section 3 describes the problem these estimators are meant to solve, and shows how that drives their design. A few of the more complicated derivations in those sections are intentionally incomplete since their purpose is to build intuitions; the reader must refer to the original papers for details. Section 4 explains the `xtabond2` and `abar` syntaxes, with examples. Section 5 concludes with a few tips for good practice.

2   Linear GMM[1]

2.1   The GMM estimator

The classic linear estimators, Ordinary Least Squares (OLS) and Two-Stage Least Squares (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; in other words, the inner products, or *moments* of the regressors with the errors are set to 0. 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 again 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 equations (moment conditions) outnumber variables (parameters), the conditions cannot be expected to hold perfectly in finite samples even if they are true asymptotically. This is the sort of problem we are interested in. To be precise, we want to fit the model:

$$y = \mathbf{x}'\beta + \varepsilon$$

$$\mathrm{E}[\mathbf{z}\varepsilon] = \mathbf{0}$$

$$\mathrm{E}[\varepsilon|\mathbf{z}] = 0$$

where β is a column of coefficients, y and ε are random variables, $\mathbf{x} = [\mathbf{x}_1 \ldots \mathbf{x}_k]'$ is a column of k regressors, $\mathbf{z} = [\mathbf{z}_1 \ldots \mathbf{z}_j]'$ is column of j instruments, $\mathbf{x}$ and $\mathbf{z}$ may share elements, and $j \geq k$. We use $\mathbf{X}$, $\mathbf{Y}$, and

---

[1]For another introduction to GMM, see Baum, Schaffer, and Stillman (2003). For a full account, see Ruud (2000, chs. 21–22). Both sources greatly influence this account.

$\mathbf{Z}$ to represent matrices of N observations for $\mathbf{x}$, y, and $\mathbf{z}$, and define $\mathbf{E} = \mathbf{Y} - \mathbf{X}\beta$. Given an estimate $\hat{\beta}$, the empirical residuals are $\hat{\mathbf{E}} = [\hat{e}_1 \dots \hat{e}_N]' = \mathbf{Y} - \mathbf{X}\hat{\beta}$. We make no assumption at this point about $\mathrm{E}[\mathbf{E}\mathbf{E}'|\mathbf{Z}] \equiv \mathbf{\Omega}$ except that it exists.

The challenge in estimating this model is that while all the instruments are theoretically orthogonal to the error term ($\mathrm{E}[\mathbf{z}\varepsilon] = \mathbf{0}$), trying to force the corresponding vector of empirical moments, $\mathrm{E}_N[\mathbf{z}\varepsilon] = \frac{1}{N}\mathbf{Z}'\hat{\mathbf{E}}$, to zero creates a system with more equations than variables if instruments outnumber parameters. The specification is *overidentified*. Since 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 $\mathrm{E}_N[\mathbf{z}\varepsilon]$.

In the Generalized Method of Moments, one defines that magnitude through a generalized metric, based on a positive semi-definite quadratic form. Let $\mathbf{A}$ be the matrix for such a quadratic form. Then the metric is:

$$\|\mathrm{E}_N[\mathbf{z}\varepsilon]\|_{\mathbf{A}} = \left\|\frac{1}{N}\mathbf{Z}'\hat{\mathbf{E}}\right\|_{\mathbf{A}} \equiv N\left(\frac{1}{N}\mathbf{Z}'\hat{\mathbf{E}}\right)' \mathbf{A}\left(\frac{1}{N}\mathbf{Z}'\hat{\mathbf{E}}\right) = \frac{1}{N}\hat{\mathbf{E}}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\hat{\mathbf{E}}. \tag{1}$$

To derive the implied GMM estimate, call it $\hat{\beta}_{\mathbf{A}}$, we solve the minimization problem $\hat{\beta}_{\mathbf{A}} = \operatorname{argmin}_{\hat{\beta}}\left\|\mathbf{Z}'\hat{\mathbf{E}}\right\|_{\mathbf{A}}$, whose solution is determined by $\mathbf{0} = \frac{d}{d\hat{\beta}}\left\|\mathbf{Z}'\hat{\mathbf{E}}\right\|_{\mathbf{A}}$. Expanding this derivative with the chain rule gives:

$$\mathbf{0} = \frac{d}{d\hat{\beta}}\left\|\mathbf{Z}'\hat{\mathbf{E}}\right\|_{\mathbf{A}} = \frac{d}{d\hat{\mathbf{E}}}\left\|\mathbf{Z}'\hat{\mathbf{E}}\right\|_{\mathbf{A}}\frac{d\hat{\mathbf{E}}}{d\hat{\beta}} = \frac{d}{d\hat{\mathbf{E}}}\left(\frac{1}{N}\hat{\mathbf{E}}'\left(\mathbf{Z}\mathbf{A}\mathbf{Z}'\right)\hat{\mathbf{E}}\right)\frac{d\left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)}{d\hat{\beta}} = \frac{2}{N}\hat{\mathbf{E}}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\left(-\mathbf{X}\right).$$

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 $\mathbf{A}$ a symmetric matrix. Dropping the factor of $-2/N$ and transposing,

$$\mathbf{0} = \hat{\mathbf{E}}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X} = \left(\mathbf{Y} - \mathbf{X}\hat{\beta}_{\mathbf{A}}\right)'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X} = \mathbf{Y}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X} - \hat{\beta}_{\mathbf{A}}'\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}$$
$$\Rightarrow \mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\hat{\beta}_{\mathbf{A}}' = \mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{Y}$$
$$\Rightarrow \hat{\beta} = \left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{Y} \tag{2}$$

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

The estimator is consistent, meaning that it converges in probability to β as sample size goes to infinity (Hansen 1982). But it is not in general unbiased, as subsection 2.6 discusses, because in finite samples the instruments are not in general perfectly uncorrelated with the endogenous components of the instrumented

regressors (correlation coefficients between finite samples of uncorrelated variables are usually not exactly 0).

For future reference, we note that the error of the estimator is the corresponding projection of the true model errors:

$$
\begin{aligned}
\hat{\beta}_{\mathbf{A}} - \beta &= \left(\mathbf{X'ZAZ'X}\right)^{-1}\mathbf{X'ZAZ'}\left(\mathbf{X}\beta + \mathbf{E}\right) - \beta \\
&= \left(\mathbf{X'ZAZ'X}\right)^{-1}\mathbf{X'ZAZ'X}\beta + \left(\mathbf{X'ZAZ'X}\right)^{-1}\mathbf{X'ZAZ'E} - \beta \\
&= \left(\mathbf{X'ZAZ'X}\right)^{-1}\mathbf{X'ZAZ'E}.
\end{aligned}
\tag{3}
$$

2.2 Efficiency

It can be seen from (2) that multiplying $\mathbf{A}$ by a non-zero scalar would not change $\hat{\beta}_{\mathbf{A}}$. But up to a factor of proportionality, each choice of $\mathbf{A}$ implies a different linear, consistent estimator of β. Which $\mathbf{A}$ should the researcher choose? Setting $\mathbf{A} = \mathbf{I}$, the identity matrix, 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 mean-zero instruments, one drawn from a variable with variance 1, the other from a variable with variance 1,000. Moments based on the second would easily dominate 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 essentially double the weight of poverty relative to other instruments. Notice that in both these examples, the inefficiency would theoretically be signaled by high variance or covariance among moments. This suggests that making A scalar is inefficient unless the moments $\frac{1}{N}\mathbf{z}_i'\mathbf{E}$ have equal variance and are uncorrelated—that is, if Var $[\mathbf{Z'E}]$ is itself scalar. This is in fact the case, as will be seen.[2]

But that negative conclusion hints at the general solution. For efficiency, $\mathbf{A}$ 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, it would efficiently down-weight one or both of the poverty proxies. In general, for efficiency, we weight by the inverse of the variance matrix of the moments:

$$
\mathbf{A}_{\text{EGMM}} = \text{Var}\left[\mathbf{Z'E}\right]^{-1} = \left(\mathbf{Z'}\,\text{Var}\left[\mathbf{E}\middle|\mathbf{Z}\right]\mathbf{Z}\right)^{-1} = \left(\mathbf{Z'\Omega Z}\right)^{-1}.
\tag{4}
$$

---

[2]This argument is identical to that for the design of Generalized Least Squares, except that GLS is derived with reference to the errors $\mathbf{E}$ where GMM is derived with reference to the moments $\mathbf{Z'E}$.

The "EGMM" stands for "efficient GMM." The EGMM estimator minimizes

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

Substituting this choice of $\mathbf{A}$ into (2) gives the direct formula for efficient GMM:

$$\hat{\beta}_{\mathrm{EGMM}} = \left(\mathbf{X}'\mathbf{Z} \left(\mathbf{Z}'\mathbf{\Omega}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{Z} \left(\mathbf{Z}'\mathbf{\Omega}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{Y} \tag{5}$$

Efficient GMM is not feasible, however, unless $\mathbf{\Omega}$ is known.

Before we move to making the estimator feasible, we demonstrate its theoretical efficiency. Let B be the vector space of linear, scalar-valued functions of the random vector $\mathbf{Y}$. This space contains all the coefficient estimates flowing from linear estimators based on $\mathbf{Y}$. For example, if $\mathbf{c} = (1\ 0\ 0\dots)$ then $\mathbf{c}\hat{\beta}_{\mathbf{A}} \in B$ is the estimated coefficient for $\mathbf{x}_1$ according to the GMM estimator implied by some $\mathbf{A}$. We define an inner product on B by $\langle b_1, b_2 \rangle = \mathrm{Cov}\left[b_1, b_2 \right]$; the corresponding metric is $\|b\|^2 = \mathrm{Var}\left[b \right]$. The assertion that (5) is efficient is equivalent to saying that for any row vector $\mathbf{c}$, the variance of the corresponding combination of coefficients from an estimate, $\left\| \mathbf{c}\hat{\beta}_{\mathbf{A}} \right\|$, is smallest when $\mathbf{A} = \mathbf{A}_{\mathrm{EGMM}}$.

In order to demonstrate that, we first show that $\left\langle \mathbf{c}\hat{\beta}_{\mathbf{A}}, \mathbf{c}\hat{\beta}_{\mathbf{A}_{EGMM}} \right\rangle$ is invariant in the choice of $\mathbf{A}$. We start with the definition of the covariance matrix and substitute in with (3) and (4):

$$
\begin{aligned}
\left\langle \mathbf{c}\hat{\beta}_{\mathbf{A}}, \mathbf{c}\hat{\beta}_{\mathbf{A}_{EGMM}} \right\rangle &= \mathrm{Cov}\left[\mathbf{c}\hat{\beta}_{\mathbf{A}}, \mathbf{c}\hat{\beta}_{\mathbf{A}_{GMM}} \right] \\
&= \mathrm{Cov}\left[\mathbf{c} \left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{Y}, \mathbf{c} \left(\mathbf{X}'\mathbf{Z}\mathbf{A}_{\mathrm{EGMM}}\mathbf{Z}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{Z}\mathbf{A}_{\mathrm{EGMM}}\mathbf{Z}'\mathbf{Y} \right] \\
&= \mathbf{c} \left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}' \mathrm{E}\left[\mathbf{E}\mathbf{E}' | \mathbf{Z} \right] \mathbf{Z} \left(\mathbf{Z}'\mathbf{\Omega}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{X} \left(\mathbf{X}'\mathbf{Z} \left(\mathbf{Z}'\mathbf{\Omega}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1} \mathbf{c}' \\
&= \mathbf{c} \left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{\Omega}\mathbf{Z} \left(\mathbf{Z}'\mathbf{\Omega}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{X} \left(\mathbf{X}'\mathbf{Z} \left(\mathbf{Z}'\mathbf{\Omega}Z \right)^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1} \mathbf{c}' \\
&= \mathbf{c} \left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X} \left(\mathbf{X}'\mathbf{Z} \left(\mathbf{Z}'\mathbf{\Omega}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1} \mathbf{c}' \\
&= \mathbf{c} \left(\mathbf{X}'\mathbf{Z} \left(\mathbf{Z}'\mathbf{\Omega}\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{X} \right)^{-1} \mathbf{c}'.
\end{aligned}
$$

This does not depend on $\mathbf{A}$. As a result, for any $\mathbf{A}$, $\left\langle \mathbf{c}\hat{\beta}_{\mathbf{A}_{EGMM}}, \mathbf{c} \left(\hat{\beta}_{\mathbf{A}_{EGMM}} - \hat{\beta}_{\mathbf{A}} \right) \right\rangle = \left\langle \mathbf{c}\hat{\beta}_{\mathbf{A}_{EGMM}}, \mathbf{c}\hat{\beta}_{\mathbf{A}_{EGMM}} \right\rangle - \left\langle \mathbf{c}\hat{\beta}_{\mathbf{A}_{EGMM}}, \mathbf{c}\hat{\beta}_{\mathbf{A}} \right\rangle = 0$. That is, the difference between any linear GMM estimator and the EGMM estimator is orthogonal to the latter. By the Pythagorean Theorem, $\left\| \mathbf{c}\hat{\beta}_{\mathbf{A}} \right\|^2 = \left\| \mathbf{c}\hat{\beta}_{\mathbf{A}} - \mathbf{c}\hat{\beta}_{\mathbf{A}_{EGMM}} \right\|^2 + \left\| \mathbf{c}\hat{\beta}_{\mathbf{A}_{EGMM}} \right\|^2 \geq \left\| \mathbf{c}\hat{\beta}_{\mathbf{A}_{EGMM}} \right\|^2$, which suffices to prove the assertion. This result is akin to the fact if there is a ball in midair, the point on the ground closest to the ball (analogous to the efficient estimator) is the one 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 efficient GMM. 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

$$\mathbf{Z}'\mathbf{Y} = \mathbf{Z}'\mathbf{X}\beta + \mathbf{Z}'\mathbf{E}, \tag{6}$$

which *is* asymptotically amenable to OLS, since the regressors, $\mathbf{Z}'\mathbf{X}$, are now orthogonal to the errors: $\mathrm{E}\left[(\mathbf{Z}'\mathbf{X})' \mathbf{Z}'\mathbf{E}\right] = (\mathbf{Z}'\mathbf{X})' \mathrm{E}\left[\mathbf{Z}'\mathbf{E}|\mathbf{Z}\right] = \mathbf{0}$ (Holtz-Eakin, Newey, and Rosen 1988). Still, though, OLS is not in general efficient on the transformed equation, since the errors are not i.i.d.—$\mathrm{Var}\left[\mathbf{Z}'\mathbf{E}\right] = \mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}$, which cannot be assumed scalar. To solve this problem, we transform the equation again:

$$(\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z})^{-1/2} \mathbf{Z}'\mathbf{Y} = (\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z})^{-1/2} \mathbf{Z}'\mathbf{X}\beta + (\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z})^{-1/2} \mathbf{Z}'\mathbf{E}. \tag{7}$$

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

$$\mathbf{Y}^* = \mathbf{X}^*\beta + \mathbf{E}^*. \tag{8}$$

Since

$$\mathrm{Var}\left[\mathbf{E}^*|\mathbf{Z}\right] = (\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z})^{-1/2} \mathbf{Z}' \mathrm{Var}\left[\mathbf{E}|\mathbf{Z}\right] \mathbf{Z} (\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z})^{-1/2} = (\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z})^{-1/2} \mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z} (\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z})^{-1/2} = \mathbf{I}.$$

this version has spherical errors. So the Gauss-Markov Theorem guarantees the efficiency of OLS on (8), which is, by definition, Generalized Least Squares on (6): $\hat{\beta}_{GLS} = \left(\mathbf{X}^{*\prime}\mathbf{X}^*\right)^{-1} \mathbf{X}^{*\prime}\mathbf{Y}^*$. Unwinding with the definitions of $\mathbf{X}^*$ and $\mathbf{Y}^*$ yields efficient GMM, just as in (5).

Efficient GMM, then, is GLS on $\mathbf{Z}$-moments. Where GLS projects $\mathbf{Y}$ into the column space of $\mathbf{X}$, 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 $\mathbf{Z}'\mathbf{Y}$, namely $\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}$, which is also the variance ellipsoid of the moments, into the column space of $\mathbf{Z}'\mathbf{X}$. If $\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}$ 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. 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 subsection—if $\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}$ is not spherical—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, then reverse the distortion.

2.3 Feasibility

Making efficient GMM practical requires a feasible estimator for the central expression, $\mathbf{Z}'\mathbf{\Omega}\mathbf{Z}$. The simplest case is when the errors (not the moments of the errors) are believed to be homoskedastic, with $\mathbf{\Omega}$ of the form $\sigma^2\mathbf{I}$. Then, the EGMM estimator simplifies to two-stage least squares (2SLS):

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

In this case, EGMM *is* 2SLS.[3] (The Stata commands `ivreg` and `ivreg2` (Baum, Schaffer, and Stillman 2003) implement 2SLS.) 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 $\hat{\mathbf{\Omega}}$ is constructed based on a formula that itself is not asymptotically convergent to $\mathbf{\Omega}$, but which has the property that $\frac{1}{N}\mathbf{Z}'\hat{\mathbf{\Omega}}\mathbf{Z}$ is a consistent estimator of $\frac{1}{N}\mathbf{Z}'\mathbf{\Omega}\mathbf{Z}$ under given assumptions. The result is the feasible efficient GMM estimator:

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

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

$$\hat{\mathbf{\Omega}} = \begin{bmatrix} \hat{e}_1^2 & & & \\ & \hat{e}_2^2 & & \\ & & \ddots & \\ & & & \hat{e}_N^2 \end{bmatrix}.$$

---

[3]However, even when the two are identical in theory, in finite samples, the feasible efficient GMM algorithm we shortly develop produces different results from 2SLS. And these are potentially inferior since two-step standard errors are often downward biased. See subsection 2.4 of this paper and Baum, Schaffer, and Stillman (2003).

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

$$
\hat{\boldsymbol{\Omega}}_i = \hat{\mathbf{E}}_i \hat{\mathbf{E}}_i' = \begin{bmatrix} \hat{e}_{i1}^2 & \hat{e}_{i1}\hat{e}_{i2} & \cdots & \hat{e}_{i1}\hat{e}_{iT} \\ \hat{e}_{i2}\hat{e}_{i1} & \hat{e}_2^2 & \cdots & \hat{e}_{i2}\hat{e}_{iT} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{e}_{iT}\hat{e}_{i1} & \cdots & \cdots & \hat{e}_{iT}^2 \end{bmatrix}.
\tag{9}
$$

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

A problem remains: where do the $\hat{e}$ 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 asymptotically *efficient*. Theoretically, any full-rank choice of $\mathbf{A}$ for the initial estimate will suffice. Usual practice is to choose $\mathbf{A} = (\mathbf{Z}'\mathbf{H}\mathbf{Z})^{-1}$, where $\mathbf{H}$ is an "estimate" of $\boldsymbol{\Omega}$ 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 $\boldsymbol{\Omega}$ in (5) with some reasonable but arbitrary $\mathbf{H}$, yielding $\hat{\beta}_1$ (one-step GMM); obtain the residuals from this estimation; use these to construct a sandwich proxy for $\boldsymbol{\Omega}$, call it $\hat{\boldsymbol{\Omega}}_{\hat{\beta}_1}$; rerun the GMM estimation, setting $\mathbf{A} = \left(\mathbf{Z}'\hat{\boldsymbol{\Omega}}_{\hat{\beta}_1}\mathbf{Z}\right)^{-1}$. This *two-step* estimator, $\hat{\beta}_2$, is asymptotically efficient and robust to whatever patterns of heteroskedasticity and cross-correlation the sandwich covariance estimator models. In sum:

$$
\begin{aligned}
\hat{\beta}_1 &= \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{H}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{H}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y} \tag{10}\\
\hat{\beta}_2 &= \hat{\beta}_{\text{EFGMM}} = \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\hat{\boldsymbol{\Omega}}_{\hat{\beta}_1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\hat{\boldsymbol{\Omega}}_{\hat{\beta}_1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y}
\end{aligned}
$$

Historically, researchers often reported one-step results as well because of downward bias in the computed standard errors in two-step. But as the next subsection explains, Windmeijer (2005) has greatly reduced this problem.

2.4 Estimating standard errors

The true variance of a linear GMM estimator is

$$
\begin{aligned}
\text{Var}\left[\hat{\beta}_{\mathbf{A}}\Big|\,\mathbf{Z}\right] &= \text{Var}\left[\left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{Y}\right] \\
&= \text{Var}\left[\beta + \left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{E}\Big|\,\mathbf{Z}\right] \\
&= \text{Var}\left[\left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{E}\Big|\,\mathbf{Z}\right] \\
&= \left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\,\text{Var}\left[\mathbf{E}|\,\mathbf{Z}\right]\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1}. \quad (11)
\end{aligned}
$$

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 weighting matrix for the instruments, discussed above, does not render the parameter estimates inconsistent even when based on incorrect assumptions about the variance of the errors, analogously substituting $\mathbf{H}$ for $\boldsymbol{\Omega}$ in (11) can make the estimate of their *variance* inconsistent. The standard error estimates will not be "robust" to heteroskedasticity or serial correlation in the errors. Fortunately, they can be made so in the usual way, replacing $\boldsymbol{\Omega}$ in (11) with a sandwich-type proxy based on the one-step residuals. This yields the feasible, robust estimator for the one-step standard errors:

$$
\widehat{\text{Var}}_{\text{r}}\left[\hat{\beta}_1\right] = \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{H}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{H}\mathbf{Z}\right)^{-1}\mathbf{Z}'\hat{\boldsymbol{\Omega}}_{\hat{\beta}_1}\mathbf{Z}\left(\mathbf{Z}'\mathbf{H}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{H}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}.
$$

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 in a sense an empirical question. The case in point: for (infeasible) efficient GMM, in which $\mathbf{A} = \left(\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}\right)^{-1}$, (11) simplifies to $\text{Var}\left[\hat{\beta}_{\mathbf{A}_{\text{EGMM}}}\right] = \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}$, a feasible, consistent estimate of which is $\widehat{\text{Var}}\left[\hat{\beta}_2\right] \equiv \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\hat{\boldsymbol{\Omega}}_{\hat{\beta}_1}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}$. 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 is 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. Since the number of moment covariances to be estimated for FEGMM, namely the distinct elements of the symmetric $\text{Var}\,[\mathbf{Z}'\mathbf{E}]$, is $j(j+1)$, these covariances can easily outstrip the statistical power of a finite sample. In fact, it is not hard for $j(j+1)$ to exceed N. When statistical power is that low, it becomes hard to distinguish means from variances. 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. The problem is analogous to that of estimating the population variances of a hundred distinct variables each with an absurdly small sample. If the samples have 1 observation each, it is impossible to estimate the variances. If they have 2 each, the sample standard deviation will tend to be half the population standard deviation, which is why small-sample corrections factors of the form $N/(N-k)$ are necessary in estimating population values.

This phenomenon does not bias coefficient estimates since identification still flows from instruments believed to be exogenous. But it can produce spurious precision in the form of implausibly good standard errors.

Windmeijer (2005) devises a small-sample correction for the two-step standard errors. The starting observation is that despite appearances in (10), $\hat{\beta}_2$ is not simply linear in the random vector $\mathbf{Y}$. It is also a function of $\hat{\mathbf{\Omega}}_{\hat{\beta}_1}$, which depends on $\hat{\beta}_1$, which depends on $\mathbf{Y}$ too. To express the full dependence of $\hat{\beta}_2$ on $\mathbf{Y}$, let

$$g\left(\mathbf{Y},\hat{\mathbf{\Omega}}\right) = \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\hat{\mathbf{\Omega}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\hat{\mathbf{\Omega}}\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{E}. \tag{12}$$

By (3), this is the error of the GMM estimator associated with $\mathbf{A} = \left(\mathbf{Z}'\hat{\mathbf{\Omega}}\mathbf{Z}\right)^{-1}$. g is infeasible since the true disturbances, $\mathbf{E}$, are unobserved. In the second step of FEGMM, where $\hat{\mathbf{\Omega}} = \hat{\mathbf{\Omega}}_{\hat{\beta}_1}$, $g\left(\mathbf{Y},\hat{\mathbf{\Omega}}_{\hat{\beta}_1}\right) = \hat{\beta}_2 - \beta$, so g has the same variance as $\hat{\beta}_2$, 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 $\hat{\beta}_2$ treats $\hat{\mathbf{\Omega}}_{\hat{\beta}_1}$ as infinitely precise. That is appropriate for one-step GMM, where $\hat{\mathbf{\Omega}} = \mathbf{H}$ is constant. But it is wrong in two-step, in which $\mathbf{Z}'\hat{\mathbf{\Omega}}_{\hat{\beta}_1}\mathbf{Z}$, the estimate of the second moments of the $\mathbf{Z}$-moments and the basis for reweighting, is imprecise. To compensate, Windmeijer develops a fuller formula for the dependence of g on the data via both its arguments, 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 $\hat{\beta}_1$, around the true (and unobserved) β:

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

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

$$g\left(\mathbf{Y},\hat{\mathbf{\Omega}}_{\hat{\beta}_1}\right) \approx g\left(\mathbf{Y},\hat{\mathbf{\Omega}}_{\beta}\right) + \mathbf{D}g\left(\mathbf{Y},\mathbf{H}\right). \tag{13}$$

Windmeijer expands the derivative in the definition of $\mathbf{D}$ using matrix calculus on (12), then replaces infeasible terms within it, such as $\hat{\mathbf{\Omega}}_{\beta}$, β, and $\mathbf{E}$, with feasible approximations. It works out that the result, $\hat{\mathbf{D}}$, is the $k \times k$ matrix whose p^{th} column is

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

where $\hat{\beta}_p$ is the pth element of $\hat{\beta}$. The formula for the $\partial\hat{\mathbf{\Omega}}_{\hat{\beta}}/\partial\hat{\beta}_p$ within this expression depends on that for $\hat{\mathbf{\Omega}}_{\hat{\beta}}$. In the case of clustered errors on a panel, $\hat{\mathbf{\Omega}}_{\hat{\beta}}$ has blocks $\hat{\mathbf{E}}_{1,i}\hat{\mathbf{E}}'_{1,i}$, so by the product rule $\partial\hat{\mathbf{\Omega}}_{\hat{\beta}}/\partial\hat{\beta}_p$ has blocks $\partial\hat{\mathbf{E}}_{1,i}/\partial\hat{\beta}_p\hat{\mathbf{E}}'_{1,i} + \hat{\mathbf{E}}_i\partial\hat{\mathbf{E}}'_{1,i}/\partial\hat{\beta}_p = -\mathbf{x}_{p,i}\hat{\mathbf{E}}'_{1,i} - \hat{\mathbf{E}}_{1,i}\mathbf{x}'_{p,i}$, where $\hat{\mathbf{E}}_{1,i}$ 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 (13), i.e., the *corrected* estimate of the variance of $\hat{\beta}_2$, works out to

$$\widehat{\text{Var}}_{\text{c}}\left[\hat{\beta}_2\right] = \widehat{\text{Var}}\left[\hat{\beta}_2\right] + \hat{\mathbf{D}}\widehat{\text{Var}}\left[\hat{\beta}_2\right] + \widehat{\text{Var}}\left[\hat{\beta}_2\right]\hat{\mathbf{D}}' + \hat{\mathbf{D}}\widehat{\text{Var}}_{\text{r}}\left[\hat{\beta}_1\right]\hat{\mathbf{D}}'$$

The first term is the uncorrected variance estimate, and the last contains the robust one-step estimate.

In difference GMM regressions on simulated panels, Windmeijer finds that the two-step efficient GMM performs somewhat better than one-step 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 robust one-step.[4]

2.5 The Sargan/Hansen test of overidentifying restrictions

A crucial assumption for the validity of GMM estimates is of course that the instruments are exogenous. If the estimation is exactly identified, detection of invalid instruments is impossible because even when

---

[4]`xtabond2` offers both.

E[$\mathbf{z}\varepsilon$] $\neq$ **0**, the estimator will choose $\hat{\beta}$ so that $\mathbf{Z}'\hat{\mathbf{E}} = \mathbf{0}$ exactly. But if the system 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 $\frac{1}{N}\mathbf{Z}'\hat{\mathbf{E}}$ is randomly distributed around **0**. A Wald test can check this hypothesis. If it holds, then

$$\left(\frac{1}{N}\mathbf{Z}'\hat{\mathbf{E}}\right)' \text{Var}\left[\frac{1}{N}\mathbf{Z}'\hat{\mathbf{E}}\right]^{-1} \frac{1}{N}\mathbf{Z}'\hat{\mathbf{E}} = \frac{1}{N}\left(\mathbf{Z}'\hat{\mathbf{E}}\right)' \mathbf{A}_{\text{EGMM}}\mathbf{Z}'\hat{\mathbf{E}} \tag{14}$$

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 $\mathbf{A}_{\text{EGMM}}$. In other words, it is just the minimized value of the criterion expression in (1) for an efficient GMM estimator. If $\mathbf{\Omega}$ is scalar, then $\mathbf{A}_{\text{EGMM}} = (\mathbf{Z}'\mathbf{Z})^{-1}$. In this case, the Hansen test coincides with the Sargan (1958) test and is consistent for "non-robust" GMM. But if non-sphericity is suspected in the errors, as in robust one-step GMM, the Sargan test statistic—$\frac{1}{N}\left(\mathbf{Z}'\hat{\mathbf{E}}\right)'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{E}}$—is inconsistent. In that case, 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 in order 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" 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 since it imposes fewer moment conditions. The difference-in-Sargan test is of course only feasible if this unrestricted regression is exactly or over-identified. (See Baum, Schaffer, and Stillman (2003).)

The Sargan/Hansen test should not be relied upon too faithfully, as it is prone to weakness. Intuitively speaking, when we apply it after GMM, we are first trying to get $\frac{1}{N}\mathbf{Z}'\hat{\mathbf{E}}$ close to 0, then testing whether it is close to 0. Counterintuitively, however, the test actually grows weaker the more instruments there are whose moments are being minimized, as the next subsection discusses.

2.6 The problem of too many instruments

The difference and system GMM estimators described in the next section can generate moment conditions prolifically, with the instrument count quadratic in time dimension, T. This can cause several problems in finite samples. First, since 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. This does not compromise the coefficient estimates (again, any choice of **A** will give a consistent estimator), but does dramatize the distance of FEGMM from the asymptotic ideal. And it can weaken the Sargan/Hansen test to the point where it generates implausibly good p values of 1.000 (Bowsher 2002). Indeed, Sargan himself (1958) determined without the aid of modern computers that the error in the 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."

In addition, 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^2's of the first-stage regressions are 1 and the second-stage results match those of (biased) OLS.

Unfortunately, there appears to be little guidance from the literature on how many instruments is "too many" (Ruud 2000, p. 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%. On the other hand, the average parameter estimate only rose from 0.9810 to 0.9866, against a true value of 1.000. `xtabond2` issues a warning when instruments outnumber individuals in the panel, as a minimally arbitrary rule of thumb. Windmeijer's finding arguably indicates that that limit 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 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:

1. The process may be dynamic, with current realizations of the dependent variable influenced by past ones.

2. There may be arbitrarily distributed fixed individual effects in the dynamic, so that the dependent variable consistently changes faster for some observational units than others. This argues against cross-section regressions, which must essentially assume fixed effects away, and in favor of a panel set-up, where variation over time can be used to identify parameters.

3. Some regressors may be endogenous.

4. The idiosyncratic disturbances (those apart from the fixed effects) may have individual-specific patterns of heteroskedasticity and serial correlation.

5. The idiosyncratic disturbances are uncorrelated across individuals.

In addition, some secondary worries shape the design:

6. Some regressors may be predetermined but not strictly exogenous: even if independent of current disturbances, still influenced by past ones. The lagged dependent variable is an example.

7. The number of time periods of available data, T, may be small. (The panel is "small T, large N.")

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

8. 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:

$$
\begin{aligned}
y_{it} &= \alpha y_{i,t-1} + \mathbf{x}'_{it}\beta + \varepsilon_{it} & (15) \\
\varepsilon_{it} &= \mu_i + v_{it} \\
\mathrm{E}\left[\mu_i\right] &= \mathrm{E}\left[v_{it}\right] = \mathrm{E}\left[\mu_i v_{it}\right] = 0
\end{aligned}
$$

14

Here the disturbance term has two orthogonal components: the fixed effects, μ_i, and the idiosyncratic shocks, v_{it}. Note that we can rewrite (15) as

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

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

In this section, we start with the classic OLS estimator applied to (15), and then modify it step by step to address all these concerns, 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 data set is based on a sample of 140 U.K. firms surveyed annually in 1976–84. The panel is unbalanced, with some firms having more observations than others. Since hiring and firing workers is costly, we expect employment to adjust with delay to changes in factors such as 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 argues for including several lags 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 data set is on the Stata web site. To download it in Stata, type `webuse abdata`.[5] The data set indexes observations by the firm identifier, `id`, and `year`. The variable `n` is firm employment, `w` is the firm's wage level, `k` is the firm's gross capital, and `ys` is aggregate output in the firm's sector, as a proxy for demand; all variables are in logarithms. Variables names ending in `L1` or `L2` indicate lagged copies. In their model, Arellano and Bond include two copies each of employment and wages (current and one-period lag) in their employment equation, three copies each of capital and sector-level output, and time dummies.

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

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

      Source |       SS       df       MS              Number of obs =     751
-------------+------------------------------           F( 16,   734) = 8136.58
       Model | 1343.31797      16  83.9573732           Prob > F      =  0.0000
    Residual | 7.57378164     734  .010318504           R-squared     =  0.9944
-------------+------------------------------           Adj R-squared =  0.9943
       Total | 1350.89175     750   1.801189            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 (15) in general, is that $y_{i,t-1}$ is endogenous to the fixed effects in the error term, which gives rise to "dynamic panel bias." To see this, consider the possibility that a firm experiences a large, negative employment shock for some reason not modeled, say in 1980, so that the shock goes into the error term. All else equal, the apparent fixed effect for that firm for the entire 1976–84 period—the deviation of its average unexplained employment from the sample average—will appear 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 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. Note that here $T = 9$. If T were large, one 1980 shock's impact 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              Number of obs =      751
-------------+------------------------------           F(155,   595) =   983.39
       Model |  1345.63898   155  8.68154179           Prob > F      =   0.0000
```

```
       Residual |  5.25277539    595   .008828194          R-squared       =  0.9961
                                                           Adj R-squared  =  0.9951
          Total |  1350.89175    750    1.801189           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

```
  .
  (remaining firm dummies 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 so transformed is the Within Groups estimator. It generates the same coefficient estimates, but standard errors that are slightly off because they do not take the pre-transformation into account[6]:

```
. 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*
       Source |       SS       df       MS              Number of obs =     751
                                                        F( 16,   734) =  180.44
        Model |  20.661288    16   1.2913305            Prob > F      =  0.0000
     Residual |  5.25277539   734    .00715637           R-squared     =  0.7973
                                                        Adj R-squared =  0.7929
        Total |  25.9140634   750   .034552084           Root MSE      =  .0846
```

n	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]

---

[6]Since `xtdata` modifies the data set, it needs to be reloaded to copy later examples.

nL1	.7329476	.0353873	20.71	0.000	.6634753	.80242
nL2	-.1394773	.0360373	-3.87	0.000	-.2102258	-.0687287
w	-.5597445	.0513496	-10.90	0.000	-.6605541	-.4589349
wL1	.3149987	.0548993	5.74	0.000	.2072204	.422777
k	.3884188	.0278697	13.94	0.000	.3337049	.4431327
kL1	-.0805185	.0346317	-2.32	0.020	-.1485076	-.0125294
kL2	-.0278013	.0295545	-0.94	0.347	-.0858227	.0302202
ys	.468666	.1108579	4.23	0.000	.2510297	.6863023
ysL1	-.6285587	.142219	-4.42	0.000	-.9077631	-.3493543
ysL2	.0579764	.1211286	0.48	0.632	-.1798234	.2957762
yr1976	(dropped)					
yr1977	(dropped)					
yr1978	(dropped)					
yr1979	.0046562	.0123816	0.38	0.707	-.0196515	.0289639
yr1980	.0112327	.0148483	0.76	0.450	-.0179175	.0403829
yr1981	-.0253693	.0195408	-1.30	0.195	-.0637318	.0129932
yr1982	-.0343973	.0201271	-1.71	0.088	-.0739109	.0051162
yr1983	-.0280344	.021675	-1.29	0.196	-.0705869	.0145181
yr1984	-.0119152	.0235643	-0.51	0.613	-.0581766	.0343461
_cons	1.79212	.4571846	3.92	0.000	.8945748	2.689665

But Within Groups does not eliminate dynamic panel bias (Nickell 1981; Bond 2002). Under the Within Groups transformation, the lagged dependent variable becomes $y_{i,t-1}^* = y_{i,t-1} - \frac{1}{T-1}(y_{i2} + ... + y_{iT})$ while the error becomes $v_{it}^* = v_{it} - \frac{1}{T-1}(v_{i2} + ... + v_{iT})$. (The use of the lagged dependent variable as a regressor restricts the sample to $t = 2, ..., T$.) The problem is that the $y_{i,t-1}$ term in $y_{i,t-1}^*$ correlates negatively with the $-\frac{1}{T-1}v_{i,t-1}$ in v_{it}^* while, symmetrically, the $-\frac{1}{T-1}y_{it}$ and v_{it} terms also move together.[7]

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 $-\frac{1}{T-1}v_{i,t-1}$ and $-\frac{1}{T-1}y_{it}$ terms above would be insignificant and the problem would disappear. 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 naive OLS regression the lagged dependent variable was positively correlated with the error, biasing its coefficient estimate upward, the opposite is the case now. Notice that 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 the range between these values—or at least near it, given that these numbers are themselves point estimates with associated confidence intervals. As Bond (2002) points out, this provides 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, 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.

---

[7]In fact, there are many other correlating term pairs, but their impact is second-order because both terms in those pairs contain a $\frac{1}{T-1}$ 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 each sum to zero. (It can be checked that 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; both are relatively canonical. 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 -1's with a diagonal of 1's just to the right. Applying the transform to (15) 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 endogenous, since the $y_{i,t-1}$ term in $\Delta y_{i,t-1} = y_{i,t-1} - y_{i,t-2}$ correlates 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, deeper lags of the regressors remain orthogonal to the error, and available as instruments.

The first-difference transform does have 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 data sets 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 since lagged observations do not enter the formula,

they are valid as instrumentss. To be precise, if w is a variable then the transform is:

$$w_{i,t+1}^{\perp} \equiv c_{it}\left(w_{it} - \frac{1}{T_{it}}\sum_{s>t} w_{is}\right). \tag{16}$$

where the sum is taken over available future observations, T_{it} is the number of such observations, and the scale factor c_{it} is $\sqrt{T_{it}/(T_{it}+1)}$. 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 identically distributed, this property too persists. In other words, $\mathbf{M}_{\perp}\mathbf{M}_{\perp}' = \mathbf{I}$.[8] 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 (16) 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 top 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

---

[8]If $\mathrm{Var}\,[v_{it}] = \mathbf{I}$ then $\mathrm{Var}\,[\mathbf{M}_{\perp}v_{it}] = \mathrm{E}\left[\mathbf{M}_{\perp}v_{it}v_{it}'\mathbf{M}_{\perp}'\right] = \mathbf{M}_{\perp}\,\mathrm{E}\left[v_{it}v_{it}'\right]\mathbf{M}_{\perp}' = \mathbf{M}_{\perp}\mathbf{M}_{\perp}'$.

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 subsection 3.5). The simplest way to incorporate either instrument is with 2SLS, which leads us to the Anderson-Hsiao (1981) "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 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 using the Stata command `ivreg`:

```
. ivreg D.n (D.nL1= nL2) D.(nL2 w wL1 k kL1 kL2 ys ysL1 ysL2 yr1979 yr1980 yr1981 yr1982 yr1983)
Instrumental variables (2SLS) regression
```

Source	SS	df	MS			Number of obs = 611
Model	-24.6768882	15	-1.64512588			F(15, 595) = 5.84
Residual	37.2768667	595	.062650196			Prob > F = 0.0000
						R-squared = .
						Adj R-squared = .
Total	12.5999785	610	.020655702			Root MSE = .2503

D.n	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
nL1						
D1.	2.307626	1.999547	1.15	0.249	-1.619403	6.234655
nL2						
D1.	-.2240271	.1814343	-1.23	0.217	-.5803566	.1323025
w						
D1.	-.8103626	.2653017	-3.05	0.002	-1.331404	-.2893209
wL1						
D1.	1.422246	1.195245	1.19	0.235	-.9251669	3.769658
k						
D1.	.2530975	.1466736	1.73	0.085	-.0349633	.5411584
kL1						
D1.	-.5524613	.6237135	-0.89	0.376	-1.777409	.6724864
kL2						
D1.	-.2126364	.2429936	-0.88	0.382	-.6898658	.264593
ys						
D1.	.9905803	.4691945	2.11	0.035	.0691015	1.912059
ysL1						
D1.	-1.937912	1.457434	-1.33	0.184	-4.800252	.9244283
ysL2						
D1.	.4870838	.5167524	0.94	0.346	-.5277967	1.501964
yr1979						
D1.	.0467148	.045459	1.03	0.305	-.0425649	.1359944
yr1980						
D1.	.0761344	.0633265	1.20	0.230	-.0482362	.2005051
yr1981						
D1.	.022623	.0564839	0.40	0.689	-.088309	.1335549
yr1982						
D1.	.0127801	.0555727	0.23	0.818	-.0963624	.1219226
yr1983						
D1.	.0099072	.0462205	0.21	0.830	-.080868	.1006824
_cons	.0159337	.0277097	0.58	0.565	-.038487	.0703545

```
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
```

21

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, and a standard error almost as large.

To improve efficiency, we can take the Anderson-Hsiao approach further, using deeper lags of the dependent variable as additional instruments. To the extent this introduces more information, it should improve efficiency. But in standard 2SLS, the deeper the lags used, the smaller the sample, since 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 a single column, as a stack of blocks like

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

The "$\cdot$" represents a missing value, which forces the deletion of that row from the data set. (Recall that 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 twice-lag of y, one for each time period, and substitute zeros for missing observations, resulting in "GMM-style" instruments:

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

(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:

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

which are based on an expectation we believe: $\mathrm{E}\left[y_{i,t-2}\varepsilon_{it}^*\right] = 0$. Alternatively, one could "collapse" this instrument set into a single column:

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

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

Having eliminated the trade-off between lag depth and sample depth, 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, since 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. In the case of $y_{i,t-1}$, 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:

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

Since in the standard, un-collapsed 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 (c.f. subsection 2.6), one can restrict the lag ranges used in generating these instrument sets. Or one can collapse them; this is non-standard but available in `xtabond2`.[9]

Although these instrument sets are part of what defines difference (and system) GMM, researchers are free to incorporate other instruments instead or in addition. 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`, then uses them in a 2SLS regression in differences. It

---

[9]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, DPD for Gauss, and performing system GMM, Calderón, Chong, and Loayza (2002) use such instruments, followed by Beck and Levine (2004) and Carkovic and Levine (2005).

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:

```
. 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)  /* replace missing with zero */
. ivreg D.n D.(nL2 w wL1 k kL1 kL2 ys ysL1 ysL2 yr1979 yr1980 yr1981 yr1982 yr1983) (D.(nL1 nL2) = z*), nocons
Instrumental variables (2SLS) regression
```

Source	SS	df	MS		Number of obs =	611
					F(15, 596) =	.
Model	8.15714895	15	.54380993		Prob > F =	.
Residual	7.29699829	596	.012243286		R-squared =	.
					Adj R-squared =	.
Total	15.4541472	611	.025293203		Root MSE =	.11065

D.n	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
nL1						
D1.	.2917489	.147383	1.98	0.048	.0022957	.5812021
nL2						
D1.	-.0653571	.0439636	-1.49	0.138	-.1516996	.0209854
D1.	(dropped)					
w						
D1.	-.5863952	.0563631	-10.40	0.000	-.6970897	-.4757008
wL1						
D1.	.2118663	.1073618	1.97	0.049	.0010128	.4227198
k						
D1.	.3876148	.0324627	11.94	0.000	.3238596	.45137
kL1						
D1.	.0735275	.0550193	1.34	0.182	-.0345277	.1815828
kL2						
D1.	.0196641	.0369952	0.53	0.595	-.0529928	.0923209
ys						
D1.	.6262124	.1178685	5.31	0.000	.3947243	.8577005
ysL1						
D1.	-.4593255	.1657888	-2.77	0.006	-.7849268	-.1337242
ysL2						
D1.	.0957105	.1304319	0.73	0.463	-.1604514	.3518725
yr1979						
D1.	.0076199	.0127743	0.60	0.551	-.0174682	.0327081
yr1980						
D1.	.021176	.01786	1.19	0.236	-.0139003	.0562522
yr1981						
D1.	-.0017659	.0228938	-0.08	0.939	-.0467283	.0431965
yr1982						
D1.	-.0165253	.0217314	-0.76	0.447	-.0592049	.0261542
yr1983						
D1.	-.0150884	.0177795	-0.85	0.396	-.0500065	.0198297

```
Instrumented:  D.nL1 D.nL2
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 z1978L2 z1979L2 z1979L3
               z1980L2 z1980L3 z1980L4 z1981L2 z1981L3 z1981L4 z1981L5
               z1982L2 z1982L3 z1982L4 z1982L5 z1982L6 z1983L2 z1983L3
               z1983L4 z1983L5 z1983L6 z1983L7 z1984L2 z1984L3 z1984L4
               z1984L5 z1984L6 z1984L7 z1984L8
```

Although this estimate is in theory not only consistent but more efficient than Anderson-Hsiao, it still seems poorly behaved. Now the coefficient estimate for lagged employment has plunged to 0.292, about 3 standard errors below the $0.733 - 1.045$ range. What is going on? As discussed in subsection 2.2, 2SLS is a good estimator under homoskedasticity. But after differencing, the disturbances v_{it}^{Δ} are far from i.i.d., far enough to greatly distort estimation. Feasible GMM directly addresses this problem, modeling the error structure more realistically, which makes it both more efficient in theory and better-behaved in practice.[10]

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 subsection 2.2 showed, FEGMM is asymptotically superior.

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

$$\text{Var}\left[\mathbf{v}_i^*\,|\,\mathbf{Z}\right] = \text{Var}\left[\mathbf{M}_*\mathbf{v}_i\,|\,\mathbf{Z}\right] = \mathbf{M}_*\,\text{Var}\left[\mathbf{v}_i\mathbf{v}_i'\,|\,\mathbf{Z}\right]\mathbf{M}_*' = \mathbf{M}_*\mathbf{M}_*'. \tag{17}$$

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

$$\begin{bmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & \ddots \\ & & \ddots & \ddots \end{bmatrix}. \tag{18}$$

As for the second FEGMM step, here we proxy $\mathbf{\Omega}^*$ with the robust, clustered estimate in (9), which is built on the assumption that errors are only correlated within individuals, not across them. For this reason, it is almost always wise to include time dummies in order to remove universal time-related shocks from the errors.

With these choices, we reach the classic Arellano-Bond (1991) difference GMM estimator for dynamic

---

[10]Apparent bias toward 0 in the coefficient estimate could also indicate weak instrumentation, a concern that motivates "System GMM," discussed later.

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 secion 4, this Stata session shows how to use the command to estimate the employment equation from before. First note that the last estimates in the previous subsection can actually be had from `xtabond2` by typing:

```
xtabond2 n L.n L2.n w L.w L(0/2).(k ys) yr*, gmmstyle(L.n) ivstyle(L2.n w L.w L(0/2).(k ys) yr*) h(1)
> noleveleq nocons 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 (18), and the results greatly improve:

```
. xtabond2 n L.n L2.n w L.w L(0/2).(k ys) yr*, gmmstyle(L.n) ivstyle(L2.n w L.w
>  L(0/2).(k ys) yr*) noleveleq nocons
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
F(16, 595)	=	42.63		avg =		4.36
Prob > F	=	0.000		max =		6

	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
n						
L1.	.2689418	.1466334	1.83	0.067	-.0190402	.5569238
L2.	-.0669834	.0437388	-1.53	0.126	-.1528845	.0189177
w						
--.	-.5723355	.0581178	-9.85	0.000	-.6864766	-.4581945
L1.	.2112242	.1050951	2.01	0.045	.0048217	.4176266
k						
--.	.3843826	.03236	11.88	0.000	.3208289	.4479363
L1.	.0796079	.0545831	1.46	0.145	-.027591	.1868069
L2.	.0231675	.0369709	0.63	0.531	-.0494419	.0957768
ys						
--.	.5976429	.1212734	4.93	0.000	.3594669	.8358189
L1.	-.4806272	.1635671	-2.94	0.003	-.8018662	-.1593882
L2.	.0581721	.1358053	0.43	0.669	-.2085439	.3248882
yr1978	.0429548	.0433702	0.99	0.322	-.0422224	.128132
yr1979	.047082	.0415766	1.13	0.258	-.0345726	.1287367
yr1980	.0566061	.0395988	1.43	0.153	-.0211644	.1343766
yr1981	.0263295	.0362365	0.73	0.468	-.0448375	.0974966
yr1982	.0018456	.0283768	0.07	0.948	-.0538852	.0575765
yr1983	-.0062288	.0197772	-0.31	0.753	-.0450704	.0326129

```
Arellano-Bond test for AR(1) in first differences: z =  -1.19  Pr > z =  0.232
Arellano-Bond test for AR(2) in first differences: z =  -0.15  Pr > z =  0.882
Sargan test of overid. restrictions: chi2(25)   =  44.85  Prob > chi2 =  0.009
Difference-in-Sargan tests of exogeneity of instrument subsets:
  Instrument group: ivstyle(L2.n w L.w L(0/2).(k ys) yr*)
    Sargan test excluding group:    chi2(10)   =  23.54  Prob > chi2 =  0.009
    Difference (null H = exogenous): chi2(15)   =  21.32  Prob > chi2 =  0.019
Warning: Sargan/Hansen tests are weak when instruments are many.
```

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).[11] Even so, the one-step coefficient on lagged employment of 0.686 (and the two-step one of 0.629) is not quite in the hoped-for 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*, gmmstyle(L.(n w k)) ivstyle(L(0/
> 2).ys yr*) noleveleq nocons robust small
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 = 90		Obs per group: min =		4
F(16, 140) = 88.07		avg =		4.36
Prob > F = 0.000		max =		6

	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf. Interval]
n					
L1.	.8179867	.0846104	9.67	0.000	.6507074 .9852659
L2.	-.1122756	.0494386	-2.27	0.025	-.2100183 -.0145329
w					
--.	-.6816685	.1403164	-4.86	0.000	-.9590816 -.4042554
L1.	.6557083	.1991534	3.29	0.001	.2619713 1.049445
k					
--.	.3525689	.1198649	2.94	0.004	.1155895 .5895483
L1.	-.1536626	.084922	-1.81	0.073	-.321558 .0142328
L2.	-.0304529	.0316251	-0.96	0.337	-.0929774 .0320715
ys					
--.	.6509498	.1865705	3.49	0.001	.2820899 1.01981
L1.	-.9162028	.2597349	-3.53	0.001	-1.429713 -.4026929
L2.	.2786584	.1825815	1.53	0.129	-.0823149 .6396318
yr1978	.0238987	.0362127	0.66	0.510	-.0476957 .0954931
yr1979	.0352258	.0346257	1.02	0.311	-.033231 .1036826
yr1980	.0502675	.035985	1.40	0.165	-.0208768 .1214119
yr1981	.0102721	.0344437	0.30	0.766	-.0578248 .0783691
yr1982	-.0111623	.0260542	-0.43	0.669	-.0626727 .0403482
yr1983	-.0069458	.0188567	-0.37	0.713	-.0442265 .030335

```
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
Hansen test of overid. restrictions: chi2(74)   =  73.72  Prob > chi2 =  0.487
Difference-in-Sargan tests of exogeneity of instrument subsets:
  Instrument group: ivstyle(L(0/2).ys yr*)
    Hansen test excluding group:     chi2(65)   =  56.99  Prob > chi2 =  0.750
    Difference (null H = exogenous): chi2(9)    =  16.72  Prob > chi2 =  1.000
Warning: Sargan/Hansen tests are weak when instruments are many.
```

---

[11]Table 4, columns (a1) and (a2).

3.4 Instrumenting with variables orthogonal to the fixed effects

Arellano and Bond compare the performance of one- and two-step difference GMM to 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 that 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 subsection 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—in symbols, that $\mathrm{E}\left[\Delta w_{it}\mu_i\right] = 0$ for all i and t. This is to say, $\mathrm{E}\left[w_{it}\mu_i\right]$ is time-invariant. If this holds, then $\Delta w_{i,t-1}$ is a valid instrument for the variables in levels:

$$\mathrm{E}\left[\Delta w_{i,t-1}\varepsilon_{it}\right] = \mathrm{E}\left[\Delta w_{i,t-1}\mu_i\right] + \mathrm{E}\left[w_{i,t-1}v_{it}\right] - \mathrm{E}\left[w_{i,t-2}v_{it}\right] = 0 + 0 - 0.$$

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, else $w_{i,t-1}$ and $w_{i,t-2}$, which may correlate with past and contemporary errors, may then correlate with future ones as well. In general, if w is endogenous, $\Delta w_{i,t-1}$ is available as an instrument since $\Delta w_{i,t-1} = w_{i,t-1} - w_{i,t-2}$ should not correlate with v_{it}; earlier realizations of Δw can instrument as well. And if w is predetermined, the contemporaneous $\Delta w_{it} = w_{it} - w_{i,t-1}$ is also valid, since $\mathrm{E}\left[w_{it}v_{it}\right] = 0$.

But the new assumption is one of stationarity, and is not trivial. Notice that the Blundell-Bond approach instruments $y_{i,t-1}$ with $\Delta y_{i,t-1}$. Lagged growth-in-y becomes an instrument in a growth-in-y regression—yet we assume that both $\Delta y_{i,t-1}$ and ε_{it} contain the fixed effect μ_i, which makes the proposition that the instrument is orthogonal to the error, that $\mathrm{E}\left[\Delta y_{i,t-1}\varepsilon_{it}\right] = 0$, counterintuitive. It can be, 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, like investment and depreciation in a Solow growth model steady state.

Blundell and Bond formalize this idea. They stipulate that α must have absolute value less than unity, so that the process is convergent. Then they derive the assumption $\mathrm{E}\left[\Delta w_{it}\mu_i\right] = 0$ 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.[12] For time-invariance of $\mathrm{E}\left[y_{it}\mu_i\right]$ to hold, the deviations of the initial observations, y_{i1}, from these long-term convergent values must not correlate with the fixed effects: $\mathrm{E}\left[\mu_i(y_{i1} - \mu_i/(1-\alpha))\right] = 0$. Otherwise, the "regression to the mean" that will occur, whereby individuals with higher initial deviations will have slower subsequent growth as they converge to the long-term level, 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 (ones with larger fixed effects) are not systematically closer or farther from their steady states than slower-growing ones.

In order to exploit the new moment conditions for the data in levels while retaining the original Arellano-Bond ones for the transformed equation, Blundell and Bond design a *system* estimator. Concretely, it involves building a stacked data set with twice the observations; in each individual's data, the transformed observations go up top, say, and the untransformed below. Formally, we produce the augmented, transformed data set by left-multiplying the original by an augmented transformation matrix,

$$
\mathbf{M}_*^+ = \left[\begin{array}{c} \mathbf{M}_* \\ \mathbf{I} \end{array} \right],
$$

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

$$
\mathbf{X}_i^+ = \left[\begin{array}{c} \mathbf{X}_i^* \\ \mathbf{X}_i \end{array} \right], \mathbf{Y}_i^+ = \left[\begin{array}{c} \mathbf{Y}_i^* \\ \mathbf{Y}_i \end{array} \right].
$$

The GMM formulas and the software still treat the system as a single-equation estimation problem since the

---

[12]This can be seen by solving $\mathrm{E}\left[y_{it}|\,\mu_i\right] = \mathrm{E}\left[y_{i,t-1}|\,\mu_i\right]$, using $y_{it} = \alpha y_{i,t-1} + \mu_i + v_{it}$.

same linear functional relationship is believed to apply in 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 coefficients estimates for other regressors. This is because all instruments for the levels equation are assumed to be orthogonal to fixed effects, thus to all time-invariant variables; in expectation, removing them from the error term therefore 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 With Groups transformation to be applied as described in subsection 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 single-column, IV-style instrument, a strictly exogenous variable w, with observation vector $\mathbf{W}$, could be transformed and entered like the regressors above,

$$\begin{bmatrix} \mathbf{W}^* \\ \mathbf{W} \end{bmatrix}, \tag{19}$$

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

$$\begin{bmatrix} \mathbf{0} \\ \mathbf{W} \end{bmatrix} \text{ and } \begin{bmatrix} \mathbf{W}^* & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{bmatrix}. \tag{20}$$

As for GMM-style instruments, the Arellano-Bond ones 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.[13] The $\underline{D}$ symbols link moments equated by the Arellano-Bond conditions on the differenced equation. The upper left one, for example, asserts $\mathrm{E}\left[w_{i1}\varepsilon_{i2}\right] = \mathrm{E}\left[w_{i1}\varepsilon_{i1}\right]$, which is equivalent to the Arellano-Bond moment condition, $\mathrm{E}\left[w_{i1}\Delta\varepsilon_{i2}\right] = 0$. The $\|L$ symbols do the same for the new Arellano-Bover conditions:

---

[13]Tue Gorgens devised these diagrams.

$$
\begin{array}{ccccccc}
\mathrm{E}\left[w_{i1}\varepsilon_{i1}\right] & \underline{\underline{D}} & \mathrm{E}\left[w_{i1}\varepsilon_{i2}\right] & \underline{\underline{D}} & \mathrm{E}\left[w_{i1}\varepsilon_{i3}\right] & \underline{\underline{D}} & \mathrm{E}\left[w_{i1}\varepsilon_{i4}\right] \\
& & \|L & & & & \\
\mathrm{E}\left[w_{i2}\varepsilon_{i1}\right] & & \mathrm{E}\left[w_{i2}\varepsilon_{i2}\right] & \underline{\underline{D}} & \mathrm{E}\left[w_{i2}\varepsilon_{i3}\right] & \underline{\underline{D}} & \mathrm{E}\left[w_{i2}\varepsilon_{i4}\right] \\
& & & & \|L & & \\
\mathrm{E}\left[w_{i3}\varepsilon_{i1}\right] & & \mathrm{E}\left[w_{i3}\varepsilon_{i2}\right] & & \mathrm{E}\left[w_{i3}\varepsilon_{i3}\right] & \underline{\underline{D}} & \mathrm{E}\left[w_{i3}\varepsilon_{i4}\right] \\
& & & & & & \|L \\
\mathrm{E}\left[w_{i4}\varepsilon_{i1}\right] & & \mathrm{E}\left[w_{i4}\varepsilon_{i2}\right] & & \mathrm{E}\left[w_{i4}\varepsilon_{i3}\right] & & \mathrm{E}\left[w_{i4}\varepsilon_{i4}\right]
\end{array}
$$

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 since the w_{it} in Δw_{it} is endogenous to the v_{it} in ε_{it}. Lagging w one period side-steps this endogeneity, yielding the valid moment expectations $\sum_i \Delta w_{i,t-1}\varepsilon_{it} = 0$ for each $t > 2$:

$$
\begin{array}{ccccccc}
\mathrm{E}\left[w_{i1}\varepsilon_{i1}\right] & & \mathrm{E}\left[w_{i1}\varepsilon_{i2}\right] & \underline{\underline{D}} & \mathrm{E}\left[w_{i1}\varepsilon_{i3}\right] & \underline{\underline{D}} & \mathrm{E}\left[w_{i1}\varepsilon_{i4}\right] \\
& & & & \|L & & \\
\mathrm{E}\left[w_{i2}\varepsilon_{i1}\right] & & \mathrm{E}\left[w_{i2}\varepsilon_{i2}\right] & & \mathrm{E}\left[w_{i2}\varepsilon_{i3}\right] & \underline{\underline{D}} & \mathrm{E}\left[w_{i2}\varepsilon_{i4}\right] \\
& & & & & & \|L \\
\mathrm{E}\left[w_{i3}\varepsilon_{i1}\right] & & \mathrm{E}\left[w_{i3}\varepsilon_{i2}\right] & & \mathrm{E}\left[w_{i3}\varepsilon_{i3}\right] & & \mathrm{E}\left[w_{i3}\varepsilon_{i4}\right] \\
& & & & & & \\
\mathrm{E}\left[w_{i4}\varepsilon_{i1}\right] & & \mathrm{E}\left[w_{i4}\varepsilon_{i2}\right] & & \mathrm{E}\left[w_{i4}\varepsilon_{i3}\right] & & \mathrm{E}\left[w_{i4}\varepsilon_{i4}\right]
\end{array}
$$

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

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

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

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, in order to minimize arbitrariness we set $\mathbf{H}$ to what $\mathrm{Var}\left[\mathbf{E}^+\right]$ would be in the simplest case. This time, however, assuming homoskedasticity with unit variance does not tie our hands enough, because the fixed effects are present in the levels errors. Consider, for example, $\mathrm{Var}\left[\varepsilon_{it}\right]$,

for some i, t, which is on the diagonal of $\text{Var}\left[\mathbf{E}^{+}\right]$. Expanding,

$$\text{Var}\left[\varepsilon_{it}\right] = \text{Var}\left[\mu_i + v_{it}\right] = \text{Var}\left[\mu_i\right] + 2\,\text{Cov}\left[\mu_i, v_{it}\right] + \text{Var}\left[v_{it}\right] = \text{Var}\left[\mu_i\right] + 0 + 1.$$

We must make an *a priori* estimate for each $\text{Var}\left[\mu_i\right]$—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

$$\text{Var}\left[\varepsilon_i^{+}\right] = \text{Var}\left[v_i^{+}\right] = \mathbf{M}_*^{+}\mathbf{M}_*^{+\prime} = \begin{bmatrix} \mathbf{M}_*\mathbf{M}_*' & \mathbf{M}_* \\ \mathbf{M}_*' & \mathbf{I} \end{bmatrix}, \tag{21}$$

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`. However, 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 2002). And the original implementation of system GMM (Blundell and Bond 1998) used $\mathbf{H} = \mathbf{I}$. These choices too are available in `xtabond2`.

For an application, Blundell and Bond return to the employment equation, using the same data set as in Arellano and Bond—and we follow suit. This time, the authors drop the deepest (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(le
> vel)) robust small
Dynamic panel-data estimation, one-step system GMM
```

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) = 1178.54				avg =		6.36
Prob > F = 0.000				max =		8

	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf. Interval]	
n						
L1.	.9356053	.0262951	35.58	0.000	.8836153	.9875953
w						
--.	-.6309761	.1180536	-5.34	0.000	-.8643889	-.3975632
L1.	.4826203	.1368872	3.53	0.001	.21197	.7532705
k						
--.	.4839299	.0538669	8.98	0.000	.3774254	.5904344
L1.	-.4243928	.0584788	-7.26	0.000	-.5400158	-.3087698
yr1977	-.0240573	.0293908	-0.82	0.414	-.082168	.0340535
yr1978	-.0176523	.0226913	-0.78	0.438	-.0625171	.0272125
yr1979	-.0026515	.0205353	-0.13	0.897	-.0432534	.0379505

```
        yr1980 │  -.0173995   .0219429    -0.79    0.429    -.0607846    .0259856
        yr1981 │  -.0435283   .0191354    -2.27    0.024    -.0813624   -.0056942
        yr1982 │  -.0096193   .0184903    -0.52    0.604    -.0461779    .0269393
        yr1983 │   .0038132   .0170186     0.22    0.823    -.0298356    .0374621
         _cons │   .5522011   .1951279     2.83    0.005     .1663985    .9380036
      ─────────────────────────────────────────────────────────────────────────
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

Hansen test of overid. restrictions: chi2(100)  = 110.70  Prob > chi2 =  0.218

Difference-in-Sargan tests of exogeneity of instrument subsets:
  Instrument group: 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 =  1.000
  Instrument group: ivstyle(yr*, equation(level))
    Hansen test excluding group:     chi2(93)  = 107.79  Prob > chi2 =  0.140
    Difference (null H = exogenous): chi2(7)   =   2.91  Prob > chi2 =  1.000
Warning: Sargan/Hansen tests are weak when instruments are many.
─────────────────────────────────────────────────────────────────────────────
```

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 (21).[14] The new point estimate of the coefficient on lagged employment is higher than that the estimate at the end of subsection 3.3, though not statistically different going by the previous standard errors. Moreover, it is within the desired range, and the reported standard error is half what it was before.

Since the additional assumption required for the validity of these estimates is nontrivial, it is worth testing. The difference-in-Sargan test in the above output for the GMM-style instruments for levels is very reassuring, with a p value of 1.000—perhaps too reassuring given the weakness of the Hansen test when instruments are numerous.

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 ε_{it}. Of course, the full disturbance v_{it} is presumed autocorrelated because it contains fixed effects, and the estimators are designed to eliminate this source of trouble. But if the ε_{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 an invalid instrument after all. The researcher would need to restrict the instrument set to lags 3 and deeper of y—unless she found order-2 serial correlation, in which case she would need to start with even deeper lags.

In order to test for autocorrelation aside from the fixed effects, the Arellano-Bond test is applied to the residuals in differences. Since 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

---

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

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 "post-determined," depending on future disturbances. (A fixed effects or Within Groups regression can violate this assumption if T is small.) Also, we will shortly see, we must assume that errors are not correlated *across* individuals. I wrote the command `abar` to make the test available after `regress`, `ivreg`, `ivreg2`, `newey`, and `newey2`.[15] So in deriving the test, we will refer to a generic GMM estimate $\hat{\beta}_{\mathbf{A}}$, applied to a dataset $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$, which may have been pre-transformed; the estimator yields residuals $\hat{\mathbf{E}}$.

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 $\frac{1}{N} \sum_i \hat{\mathbf{E}}_i^{-l} \hat{\mathbf{E}}_i$, which is zero in expectation under the null of no order-l serial correlation. Assuming errors are uncorrelated across individuals, the terms of this average are also uncorrelated and, under suitable regularity conditions, the central limit theorem assures that

$$\sqrt{N} \frac{1}{N} \sum_i \hat{\mathbf{E}}_i^{-l} \hat{\mathbf{E}}_i = \frac{1}{\sqrt{N}} \hat{\mathbf{E}}^{-l} \hat{\mathbf{E}} \tag{22}$$

is asymptotically normally distribution. Notice how this statistic is constructed on the assumption that N is large but T may not be.

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

---

[15]I also wrote `newey2`; it makes Newey-West autocorrelation-robust standard errors available for 2SLS regressions. `ivreg2` (Baum, Schaffer, and Stillman 2003) now includes this functionality.

(22) gives

$$
\frac{1}{\sqrt{N}}\hat{\mathbf{E}}^{-l\prime}\hat{\mathbf{E}} = \frac{1}{\sqrt{N}}\left(\mathbf{E}^{-l} - \mathbf{X}_i^{-l}\left(\hat{\beta}_{\mathbf{A}} - \beta\right)\right)'\left(\mathbf{E} - \mathbf{X}\left(\hat{\beta}_{\mathbf{A}} - \beta\right)\right)
$$

$$
= \frac{1}{\sqrt{N}}\mathbf{E}^{-l\prime}\mathbf{E} - \frac{\mathbf{E}^{-l\prime}\mathbf{X}}{N}\sqrt{N}\left(\hat{\beta}_{\mathbf{A}} - \beta\right)
$$

$$
-\sqrt{N}\left(\hat{\beta}_{\mathbf{A}} - \beta\right)'\frac{\mathbf{X}^{-l\prime}\mathbf{E}}{N} + \sqrt{N}\left(\hat{\beta}_{\mathbf{A}} - \beta\right)'\frac{1}{\sqrt{N}}\frac{\mathbf{X}^{-l\prime}\mathbf{X}}{N}\sqrt{N}\left(\hat{\beta}_{\mathbf{A}} - \beta\right)
$$

The last two terms drop out as $N \to \infty$. Why? Since $\hat{\beta}_{\mathbf{A}}$ is a $\sqrt{N}$-consistent estimate of β (Ruud 2000, p. 546), the $\sqrt{N}\left(\hat{\beta}_{\mathbf{A}} - \beta\right)$ terms neither diverge nor converge to 0. Meanwhile, assuming $\mathbf{x}$ is not post-determined, $\mathbf{X}^{-l\prime}\mathbf{E}/N$ goes to 0, which eliminates the third term. Finally, assuming that $\mathbf{X}^{-l\prime}\mathbf{X}/N$ does not diverge, the fourth term goes to zero. If we then substitute with (3) into the second term, the expression can be seen to converge to $\frac{1}{\sqrt{N}}\left[\mathbf{E}^{-l\prime}\mathbf{E} - \mathbf{E}^{-l\prime}\mathbf{X}\left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{E}\right]$, whose variance is consistently estimated by

$$
\frac{1}{\sqrt{N}}\left[\hat{\mathbf{E}}^{-l\prime}\widehat{\mathrm{Var}}\left[\hat{\mathbf{E}}\middle|\mathbf{Z}\right]\hat{\mathbf{E}}^{-l} - 2\hat{\mathbf{E}}^{-l\prime}\mathbf{X}\left(\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\mathbf{A}\mathbf{Z}'\widehat{\mathrm{Var}}\left[\hat{\mathbf{E}}\middle|\mathbf{Z}\right]\hat{\mathbf{E}}^{-l} + \hat{\mathbf{E}}^{-l\prime}\mathbf{X}\widehat{\mathrm{Var}}\left[\hat{\beta}_{\mathbf{A}}\right]\mathbf{X}'\hat{\mathbf{E}}^{-l}\right].
$$

(Arellano and Bond 1991). Dividing this value into (22) 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. $\hat{\mathbf{E}}^{-l}$ contains lagged, differenced errors, with observations for the levels data zeroed out in system GMM since they are not the basis for the test. $\mathbf{X}$ and $\mathbf{Z}$ hold the transformed and, in system GMM, augmented data set used in the estimation. In one-step, non-robust estimation, $\widehat{\mathrm{Var}}\left[\hat{\mathbf{E}}\middle|\mathbf{Z}\right]$ is $\hat{\sigma}^2\mathbf{H}$, where $\hat{\sigma}$ is a consistent estimate of the standard deviation of the errors in levels. Otherwise, $\hat{\mathbf{\Omega}}_{\hat{\beta}_1}$ is substituted. $\widehat{\mathrm{Var}}\left[\hat{\beta}_{\mathbf{A}}\right]$ is set to the reported variance matrix—robust or not, Windmeijer-corrected or not.[16]

There are two important lessons here for the researcher. The first is another reminder of the importance of time dummies to preventing the most likely form of cross-individual correlation, contemporaneous correlation. The test assumes no correlations in errors across individuals. Second 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

---

[16]In addition, in one-step, non-robust estimation in orthogonal deviations, the second $\widehat{\mathrm{Var}}\left[\hat{\mathbf{E}}\middle|\mathbf{Z}\right]$ is actually set to $\mathbf{M}_{\perp}\mathbf{M}_{\Delta}'$ in "difference" GMM and $\mathbf{M}_{\perp}^{+}\mathbf{M}_{\Delta}^{+\prime}$ in system GMM.

that their test has greater power than the Sargan and Hansen tests to detect lagged instruments being made invalid through autocorrelation. The test does break down, however, 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 the difference GMM estimator was in the package DPD, written in the Gauss programming language (Arellano and Bond 1988). An update, DPD98, incorporates system GMM. DPD has also been implemented in the Ox language (Doornik, Arellano, and Bond 2002). In 2001, Stata Corporation shipped `xtabond` in Stata 7. It performs 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 system GMM and the Windmeijer correction as well as other options. It was and is compatible with Stata versions 7 and later. Unlike `xtabond`, which has computationally intensive sections precompiled, `xtabond2` was written purely in Stata's interpreted ado language, which made it slow. I also wrote `abar` to make the Arellano-Bond autocorrelation test available after other estimation commands. In late 2005, I implemented `xtabond2` afresh in the Mata language shipped with Stata 9; the new version runs much faster. The latest version of the Mata code at this writing introduces a new feature, the orthogonal deviations transform. The two implementations are bundled together and the ado version automatically runs if Mata is not available.[17]

The syntax for `xtabond2` is:

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

where `gmmopt` is

```
gmmstyle(varlist [, laglimits(# #) collapse equation({diff | level | both}) passthru])
```

and `ivopt` is

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

Items in [brackets] are optional. Underlining indicates minimum allowed abbreviations. {Braces} enclose lists of choices. Options after the comma may appear in any order. All *varlist*'s can include time-series

---

[17]The Mata code requires Stata 9.1 or later. Version 9.0 users will be prompted to make a free upgrade.

operators such as `L.` and wildcard expressions such as `_I*`.

The `if` and `in` clauses are standard ones that restrict the sample. But they do not restrict the sample from which lagged variables are drawn for instrument construction. The `level`, `noconstant`, `small`, and `robust` options are also mostly standard. `level` controls the size of the reported confidence intervals, the default being 95 (percent). `small` requests small-sample corrections to the covariance matrix estimate, resulting in t instead of z test statistics for the coefficients and an F instead of Wald χ^2 test for overall fit. `noconstant` excludes the constant term from X and Z. However, it has no effect in difference GMM since differencing eliminates the constant anyway.[18] In one-step GMM, `xtabond2`'s `robust` 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 theoretically robust, `robust` triggers the Windmeijer correction.

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 GMM, one-step being the default. `noleveleq` invokes difference instead of system GMM, the default. `nodiffsargan` prevents reporting of certain difference-in-Sargan statistics (described below), which are computationally intensive since they involve re-estimating the model for each test. It has effect only in the Mata implementation, since the ado one does not perform the test. `orthogonal`, also only meaningful 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 only applies to system GMM, and only makes sense in the rare case where it is believed that there are no fixed effects that need to be purged by differencing. The `h()` option, which most users can also safely ignore, controls the choice of **H**. `h(1)` sets $\mathbf{H} = \mathbf{I}$, for both difference and system GMM. For difference GMM, `h(2)` and `h(3)` coincide, equaling the matrix in (17). They differ for system GMM, however, with `h(2)` imitating DPD for Ox and `h(3)` being the `xtabond2` default, according to (21) (see the end of subsection 3.4).

The most important thing to understand about the `xtabond2` syntax is that unlike most Stata estimation

---

[18]Here, `xtabond2` differs from `xtabond` and DPD, which normally 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, under the difference transform, it 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, in 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 in practice. Usually difference and system GMM regressions include time dummies. Since 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 variables as a group enter transformed or not.

commands, including `xtabond`, the variable list before the comma *communicates no identification information.* The first variable defines $\mathbf{Y}$ and the remaining ones define $\mathbf{X}$. None of them say anything about $\mathbf{Z}$ even though $\mathbf{X}$ and $\mathbf{Z}$ may share columns. Designing the instrument matrix is the job of the `ivstyle()` and `gmmstyle()` options after the comma, each of which may be listed multiple times, or not at all. (`noconstant` also affects $\mathbf{Z}$ in system GMM.) As a result, most regressors appear twice in a command line, once before the comma for inclusion in $\mathbf{X}$, once after as a source of IV- or GMM-style instruments. Variables that only serve as 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, 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 `iv(w1 w2, passthru)`. `ivstyle()` also generates one column per variable in system GMM, following (19). The patterns in (20) can be requested using the `equation` suboption, as in: `iv(w1 w2, eq(level))` and the compound `iv(w1 w2, eq(diff)) iv(w1 w2, eq(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 they enter $\mathbf{Z}$. 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, `gmm(w, laglimits(2 .))` specifies lags 2 and deeper 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.[19] Since 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 `gmm(w1) gmm(w2, lag(2 .))`, one could simply type `gmm(w1 L.w2)`. In all of these instances, the suboption `collapse` is available to "collapse" the instrument sets as described in subsections 3.2 and 3.4.

---

[19]If $a <= b < 0$ 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 $a <= 0 <= b$ or $b <= 0 <= a$, the contemporaneous difference is used. Tue Gorgens developed these decision rules.

gmmstyle() also has `equation()` and `passthru` suboptions, which work much like their `ivstyle()` counterparts. The exception is that `eq(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.[20] `passthru` prevents the usual differencing of instruments for the levels equation. As with `arlevels`, this produces invalid results under the usual assumptions.

The Mata version of `xtabond2` also 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, perm` at the Stata prompt), the Mata code builds a complete internal representation of $\mathbf{Z}$.[21] 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 takes 8 bytes in Mata, for a total of roughly 1.5 megabytes. Larger panels can exceed a computer's physical memory and actually even slow Mata down as the operating system is forced to repeatedly cache parts of $\mathbf{Z}$ to the hard drive, 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 since the ado version also builds a full representation of $\mathbf{Z}$, the Mata code in `space` mode still has the advantage of conserving memory.

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 different samples simultaneously. `xtabond2` reports the size of the transformed sample after difference GMM and the untransformed one after system GMM.

The Mata implementation carries out certain difference-in-Sargan tests unless `nodiffsargan` is specified. In particular, it reports a difference-in-Sargan test for each instrument group defined by an `ivtyle()` or `gmmstyle()` option, when feasible. So a clause like `gmm(x y)` implicitly requests a single test for this entire instrument group while `gmm(x) gmm(y)` requests the same estimates, but two more targeted difference-in-Sargan tests. In addition, in system GMM, the Mata version tests the GMM-style instruments for the levels equation that characterize system GMM, as a group.[22]

---

[20] Since an ordinary `gmm(w, laglim(a b))` 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 `gmm(w, laglim(a b) eq(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–b.

[21] Despite the speed setting, there is a delay the first time the Mata version of `xtabond2` runs in a Stata session, as Stata loads the function library.

[22] The reported differences-in-Sargan will generally not match what would be obtained by manually running the estimation with and without the suspect instruments. Recall from subsection 2.3 that in the full, restricted regression, the moment weighting matrix is the inverse of the estimated covariance of the moments, call it $\hat{\mathbf{S}}$, which is $\mathbf{Z}'\mathbf{HZ}$ in one-step and $\mathbf{Z}'\hat{\mathbf{\Omega}}_{\hat{\beta}_1}\mathbf{Z}$

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 algorithm and tolerances.[23]

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

`predict [type] newvarname [if exp] [in range] [, 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 $\mathbf{X}\hat{\beta}$ where $\hat{\beta}$ 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 $\mathbf{Y}$. Putting `residuals` in the command line requests $\mathbf{Y} - \mathbf{X}\hat{\beta}$, where the $\mathbf{X}\hat{\beta}$ again will be adjusted. The `difference` option requests predictions and residuals in differences.

The syntax for the post-estimation command `abar` is

`abar [if exp] [in range] [, lags(#)]`

The `lags()` option works like `xtabond2`'s `artests()` option except that it defaults to 1. `abar` can run after `regress`, `ivreg`, `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 estimated by

`xi: xtabond2 y L.y i.t, gmm(L.y) iv(i.t) robust noleveleq`

---

in two-step. In the unrestricted regressions carried out for testing purposes, `xtabond2` weights using the submatrix of the restricted $\hat{\mathbf{S}}$ corresponding to the non-suspect instruments. This reduces the chance of a negative test statistic (Baum, Schaffer, and Stillman 2003, p. 18, citing Hayashi 2000). As described in subsection 2.6, adding instruments weakens the Sargan/Hansen test and can actually reduce the statistic, which is what makes negative differences-in-Sargan more likely if the unrestricted regression is fully re-estimated.

[23]In addition, the Mata version will not perfectly handle strange and unusual expressions like `gmm(L.x, lag(-1 -1))`. This is the same as `gmm(x, lag(0 0))` in principle. But the Mata code will interpret it by lagging x, thus losing the observation of x for $t = T$, 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, gmm(L.y w2 L.w3) iv(i.t w1) two robust small orthog
```

would estimate the model with the standard choices of instruments—in this case with two-step system GMM, corrected standard errors, small-sample adjustments, and orthogonal deviations.

If the user runs system GMM without declaring instruments that are non-zero 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 data set:

```
regress n w k
abar
xtabond2 n w k, iv(w k, eq(level)) small arlevels artests(1)

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

ivreg2 n cap (w = k ys), cluster(id) gmm
abar
xtabond2 n w cap, iv(cap k ys, eq(lev)) two artests(1) arlevels
```

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

`xtabond2` can replicate results from comparable packages. To reproduce results from `xtabond`, one must usually generate a variable that `xtabond2` will difference *into* a constant, in order to mimic the way `xtabond` enters a constant directly into the difference equation. Thus, another matching pair:

```
gen cons = year
xtabond n, lags(1) pre(w, lagstruct(1,.)) pre(k, endog) robust
xtabond2 n L.n w L.w k cons, gmm(L.(w n k)) iv(cons) noleveleq robust
```

To exactly match difference GMM results from DPD for Gauss and Ox, one must also create variables that become time dummies after transformation. This example exactly imitates the regression for column (a1), Table 4 in Arellano and Bond (1991):

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

For system GMM, these gymnastics are unnecessary since DPD enters the constant and time dummies directly into the levels equation, not the difference one. This command exactly reproduces a version of the of Blundell and Bond's regression 4, Table 4, included in a demonstration file included with DPD for Ox[24]:

```
xtabond2 n L.n L(0/1).(w k) yr1978-yr1984, gmm(L.(w k n)) iv(yr1978-yr1984, eq(level)) h(2) robust two small
```

More replications from the regressions in the Arellano-Bond and Blundell-Bond papers are in two ancillary files that come with `xtabond2`, abest.do and bbest.do. In addition, greene.do reproduces a run in Greene (2002, p. 554).[25]

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.

1. *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 Arellano-Bond autocorrelation test may become unreliable.

2. *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.

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

4. *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 a single column (in `xtabond2`, with `iv(w)`). If `w` is predetermined by not strictly exogenous, standard treatment is to use lags 1 and deeper, GMM-style (`gmm(w)`). And if `w` is endogenous, standard treatment is lags 2 and deeper (`gmm(L.w)`).

5. *Mind and report the instrument count.* As discussed in subsection 2.6, too many instruments can overfit endogenous variables and fail to expunge their endogenous components. Ironically, it also weakens the power of the Hansen test to detect overidentification. Because the risk is high with these estimators,

---

[24]In the command file bbest.ox.
[25]To download them into your current directory, type `net get xtabond2` in Stata.

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. One rule of thumb is to worry if the instrument count exceeds N. If it does, then test for robustness to reducing it. Options include limiting the lags used in GMM-style instruments and, in `xtabond2`, collapsing instruments.

6. *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 in a kind of steady-state, in the sense that deviations from long-term values, controlling for covariates, are not systematically related to fixed effects.

7. *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 deviations; one- or two-step estimation; non-robust, robust, or corrected errors; and the choice of instrumenting variables and lags used.

References

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

[2] 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: 277–97.

[3] –. 1998. Dynamic Panel data estimation using DPD98 for Gauss: A guide for users.

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

[5] Baum, C., M. Schaffer, and S. Stillman. 2003. Instrumental variables and GMM: Estimation and testing. *Stata Journal* 3(1): 1-31.

[6] Beck, T., and R. Levine. 2004. Stock Markets, Banks, and Growth: Panel Evidence. *Journal of Banking and Finance* 28(3): 423–42.

[7] Blundell, R., and S. Bond. 1998. Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87: 11–143.

[8] Bond, S. 2002. Dynamic panel data models: A guide to micro data methods and practice. Working Paper 09/02. Institute for Fiscal Studies. London.

[9] Bowsher, C.G. 2002. On testing overidentifying restrictions in dynamic panel data models. *Economics Letters* 77: 211–20.

[10] Calderón, C.A., A. Chong, and N.V. Loayza. 2002. Determinants of Current Account Deficits in Developing Countries. *Contributions to Macroeconomics* 2(1): Article 2.

[11] Carkovic, M., and R. Levine. 2005. Does Foreign Direct Investment Accelerate Economic Growth? in T.H. Moran, E.M. Graham, and M. Blomström. Does Foreign Direct Investment Promote Development? Washington, DC: Institute for International Economics and Center for Global Development.

[12] Doornik, J.A., M. Arellano, and S. Bond. 2002, Panel data estimation using DPD for Ox.

[13] Greene, W.H. 2002. *Econometric Analysis*, 5th ed. Upper Saddle River, NJ: Prentice-Hall.

[14] Hansen, L. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50(3): 1029–54.

[15] Hayashi, F. 2000. *Econometrics.* 1st ed. Princeton, NJ: Princeton University Press.

[16] Holtz-Eakin, D., W. Newey, and H.S. Rosen. 1988. Estimating vector autoregressions with panel data. *Econometrica* 56: 1371–95.

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

[18] Judson, R.A., and A.L. Owen. 1999. Estimating Dynamic Panel Models: A Pracical Guide for Macroeconomists *Economics Letters* 65: 9–15.

[19] Ruud, P.A., 2000. *Classical Econometrics.* New York: Oxford University Press.

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

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