EC327: Advanced Econometrics, Spring 2007

Wooldridge, Introductory Econometrics (3rd ed, 2006)

Chapter 14:
Advanced panel data methods

Fixed effects estimators

We discussed the first difference (FD) model as one solution to the problem of unobserved heterogeneity in the context of panel data. It is not the only solution; the leading alternative is the *fixed effects* model, which will be a better solution under certain assumptions.

For a model with a single explanatory variable,

$$y_{it} = \beta_1 x_{it} + a_i + u_{it} \qquad (1)$$

If we average this equation over time for each unit i, we have

$$\bar{y}_{it} = \beta_1 \bar{x}_{it} + \bar{a}_i + \bar{u}_{it} \tag{2}$$

Subtracting the second equation from the first, we arrive at

$$y_{it} - \bar{y}_{it} = \beta_1 (x_{it} - \bar{x}_{it}) + (u_{it} - \bar{u}_{it}) \tag{3}$$

defining the *demeaned data* on $[y, x]$ as the observations of each panel with their mean values per individual removed. This algebra is known as the *within transformation*, and the estimator we derive is known as the *within estimator*. Just as OLS in a cross-sectional context only "explains" the deviations of y from its mean $\bar{y}$, the within estimator's explanatory value is derived from the comovements of y around its individual-specific mean with x around its individual-specific mean. Thus, it matters not if a unit has consistently high or low values of y and x. All that matters is how the variations around those mean values are correlated.

Just as in the case of the FD estimator, the within estimator will be unbiased and consistent if the explanatory variables are strictly exogenous: independent of the distribution of u in terms of their past, present and future values. Correlation with a_i is allowable since that term will be removed in the within transformation.

The within estimator is implemented by Stata's command `xtreg, fe` (`fe` for fixed effects). The within transformation implements what has often been called the LSDV (least squares dummy variable) model because the regression on demeaned data yields the same results as estimating the model from the original data and a set of $(N-1)$ indicator variables for all but one of the panel units. It is often not workable to estimate that LSDV model directly because we may have hundreds or thousands of individual panel units in our dataset. We can always perform the within transformation for any number

of units, though, and implement the FE model. Note that the degrees of freedom for such a model will take account of the N means that were estimated, one for each individual. Thus, unlike pooled OLS where the number of degrees of freedom would be $(NT - k)$, the degrees of freedom for the FE estimator will be $(N(T - 1) - k)$. In Stata's implementation in `xtreg, fe`, a constant term is included and a F-test is provided for the null hypothesis that all coefficients $a_i^{'}$ are zero, where $a_i^{'}$ are deviations from the mean value $\bar{a}_i$.

Analogous to the FD model, we cannot include time-invariant variables in the FE model, since the demeaning process will cause their value to be zero for all time periods. We can interact such variables with time-varying variables, though. We could interact a gender indicator with time dummies, which would allow us to estimate how the effect of gender has *changed*

over the time periods. We cannot estimate the effect of gender in the base period, though, since that is subsumed in the a_i term.

If we introduce a full set of $(T-1)$ time dummies (one for each period but the first), we cannot include any explanatory variables that have a constant *difference* over time for each individual: e.g., age in an annual dataset. The same would be true if we introduced a linear time trend rather than time dummies: it absorbs all time-constant effects.

Fixed effects or first differencing?

Two competing methods: first differencing and fixed effects. Which should we use? If T=2, it does not matter, since FD and FE methods are identical in that case. When T $\geq$ 3, the two methods do not yield the same

results, but they are both unbiased estimators of the underlying coefficient vector. Both are consistent with T fixed as $N \to \infty$. For large N and small T (a common setup in many datasets) we might be concerned with relative efficiency. When the u_{it} are serially uncorrelated (given that they are homoskedastic, this amounts to saying they are $i.i.d.$) FE will be more efficient than FD, and the standard errors reported from FE are valid. We often may assume serially uncorrelated errors, but there is no reason why that condition will necessarily hold in the data. If u_{it} follows a random walk process, then its differences will be uncorrelated, and first differencing will be the appropriate estimator. But we may often encounter an error process with some serial correlation, but not necessarily a random walk process.

When T is large and N is not very large (for instance, when we have many time periods

of data on each of a small number of units) we must be careful in using the FE estimator, since its large-sample justification relies on $N \to \infty$, not T. If FE and FD give substantively different results, it might be very hard to choose between them, and we might want to report them both.

One consideration arises when we are using an unbalanced panel—especially one in which the missing observations on some units do not appear at the beginning or end of their time series, but create "gaps" in the time series. The FE estimator has no problem with this, but the FD estimator will lose two observations when there is a single period missing in the sequence of observations for that unit. One thing we must consider is why the data are missing. If they can be considered "missing at random", this may not be problematic, but if there is some pattern to missingness we must be concerned about it.

One issue that often arises with individuals or firms is attrition: units leaving the sample. Individuals can die; firms can liquidate or be taken over. Are these events related to the variables we are using in the regression model? If so, we may want to worry about the *sample selection* problem that this entails. Nevertheless, one advantage of fixed effects is that it allows the attrition to be correlated with a_i, the unobserved fixed effect.

Two-way fixed effects

Stata lacks a command to estimate two-way fixed effects models. If the number of time periods is reasonably small, you may estimate a two-way FE model by creating a set of time indicator variables and including all but one in the regression.

The joint test that all of the coefficients on those indicator variables are zero will be a test

of the significance of time fixed effects. Just as the individual fixed effects (LSDV) model requires regressors' variation over time within each *unit*, a time fixed effect (implemented with a time indicator variable) requires regressors' variation over units within each *time period*. If we are estimating an equation from individual or firm microdata, this implies that we cannot include a "macro factor" such as the rate of GDP growth or price inflation in a model with time fixed effects, since those factors do not vary across individuals.

Random effects models

As an alternative to the individual fixed effects model, we may consider a *random effects* formulation. For a single explanatory variable, this becomes

$$y_{it} = \beta_0 + \beta_1 x_{it} + [a_i + u_{it}] \qquad (4)$$

where we explicitly include an intercept so that we can make the assumption that the unobserved effect, a_i, has a zero mean. The bracketed term is the *composite error term*, which is now assumed to have an individual-specific component and an idiosyncratic component.

In the fixed effects formulation, a_i is treated as an unknown "nuisance parameter" which, if ignored, causes bias and inconsistency in our estimators because it is correlated with one or more of the regressors. In the fixed effects *within transformation* we get rid of a_i. But if we can assume that a_i is a random variable distributed independently of x (or generally independent of all of the regressors), we can derive a more efficient estimator of the problem than fixed effects.

The *random effects* model then proceeds by using the form of the composite error term

in an optimal manner. For instance, with the assumption of independence, we could use a single cross section to optimally estimate the β vector; there would be no need for panel data. That would discard information from other cross sections, so we might rather want to use pooled OLS, which will be consistent in this case. But pooled OLS will not make optimal use of the assumed structure in the error term. The composite error term assumes that the errors arise for two reasons: one of them common to all observations on a single individual, the other purely idiosyncratic.

If we define the composite error term $v_{it} = a_i + u_{it}$ as the regression error, the v_{it} series must be serially correlated. Under the random effects assumptions,

$$corr(v_{it}, v_{is}) = \frac{\sigma_a^2}{(\sigma_a^2 + \sigma_u^2)} \; \forall \; t \neq s \qquad (5)$$

and this correlation will be positive whenever σ_a^2 is nontrivial. Indeed, in a case where a large fraction of the variation in the composite error term is due to the individual-specific component, this correlation can be quite substantial. Since the standard errors of a pooled OLS model ignore it, they will be biased.

Just as in other cases where we can explicitly model the form of serial correlation in the error process, we can use generalized least squares (GLS) to solve this problem. Although the algebra to derive the GLS estimator is quite complex, the GLS transformation itself is simple. If we define

$$\lambda = 1 - \sqrt{\left(\frac{\sigma_u^2}{(\sigma_u^2 + T\sigma_a^2)}\right)} \tag{6}$$

we will generate a weight λ, $0 \leq \lambda \leq 1$. The transformed equation is then defined in the

quasi-demeaned y and x variables:

$$(y_{it} - \lambda\bar{y}_t) = \beta_0(1-\lambda) + \beta_1(x_{it} - \lambda\bar{x}_t) + (v_{it} - \lambda\bar{v}_t) \tag{7}$$

where the overbar denotes the averages over time for each panel unit, just as in the fixed effects model.

The fixed effects (FE) model arbitrarily sets $\lambda = 1$ and fully demeans the data. As we can see from equation (**??**), that would be appropriate in this context iff the idiosyncratic error variance was very small relative to the variance of the individual effect a_i. In other cases, where both sources of variance are non-negligible, the optimal λ will be less than one. We can also consider the pooled OLS model in this context; it corresponds to a $\lambda = 0$ in which we do not transform the data at all. Arbitrarily setting $\lambda = 1$ à la FE leads to a consistent estimator of the equation, but it is

inefficient relative to the RE alternative. Because the FE model is equivalent to the LSDV formulation, it involves the loss of N degrees of freedom. Given that the a_i may be considered as nuisance parameters, if we do not care about their values, we might rather apply RE and substantially reduce the degrees of freedom lost in estimation: especially important if T is small.

We do not know λ, of course, so we must consistently estimate it. The ability to do so involves the crucial assumption that $cov(x_{it}, a_i) = 0$: the unobservable individual effects must be independently distributed of the regressors. If our estimate of λ is close to zero, the RE estimates will be similar to those of a pooled OLS model. If our estimate of λ is close to one, the RE estimates will be similar to those of a FE model. The RE estimator may be chosen in Stata by giving the command `xtreg` *depvar*

indepvars, `re`. The estimated results will display an estimate of λ. In the example of a wage equation given in the textbook (14.4), a $\widehat{\lambda} = 0.643$ is displayed, indicating that the RE estimates are likely to differ considerably from both pooled OLS and FE counterparts.

One interesting feature of the random effects estimator, evident in that example: since it involves quasi-demeaning of the data, a variable without time variation within the individual may be included in the model. Thus, if we are estimating a wage equation, we can include gender or race in a RE model, whereas it cannot be included in a FE model. However, we must ensure than any such variable satisfies the assumption that $cov(x_{it}, a_i) = 0$.

Random effects or fixed effects?

To justify RE, the necessary assumption that an individual effect can be considered independent of all regressors is often problematic. If

we are interested in testing the effect of a time-invariant variable, RE can yield such an estimate, but we should include all available time-invariant variables as controls to try to ensure that the independence assumption is satisfied.

If we are interested in evaluating the effect of a time-varying explanatory variable, can we justify the use of RE? Yes, but in realistic terms probably only in the case where the key variable is set randomly. For instance, if students are assigned randomly to sections of a course or home rooms in a K-12 context, RE would be appropriate given that the assignment variable would not be correlated with unobservables such as aptitude. On the other hand, if students are grouped by ability or test scores and assigned to home rooms accordingly, the assignment variable will not be independent of the unobservable individual aptitude, and RE will be inconsistent.

We can formally evaluate the appropriateness of the RE estimator in a given context with a *Hausman test*. A Hausman test compares the coefficient vectors from two estimators. If they are both consistent estimators, then their point estimates should not differ greatly, whereas if one of the estimators is inconsistent, its point estimates are likely to differ widely from those of a consistent estimator. In the current context, the FE estimator is always consistent, but inefficient under the null hypothesis that $cov(x_{it}, a_i) = 0$. RE is both consistent and relatively efficient under that null hypothesis, but inconsistent under the alternative. To evaluate the null hypothesis, we give the commands

```
xtreg depvar indepvars1, fe
estimates store fe
xtreg depvar indepvars2, re
estimates store re
hausman  fe re, sigmamore
```

where we note that *indepvars1* may not contain all of the regressors in *indepvars2* because the RE estimator may also estimate time-invariant effects. It is crucial that the two sets of estimates' names be given in the order shown, with the always-consistent estimator first in the `hausman` command.

The null hypothesis for the Hausman test is that RE is consistent and should be preferred. If we reject that null, RE is inappropriate and FE should be used instead. However, like many tests, the Hausman test is performed conditional on proper specification of the underlying model. If we have omitted an important explanatory variable from both forms of the model, then we are comparing two inconsistent estimators of the population model. When a rejection is received, specification tests should be used to try to rule out this possibility.

We might consider RE as more appropriate when applied to a random sample of individuals (such as a sample of workers, or the unemployed, or those who have completed a job training program), and FE the better choice when we consider observations corresponding to a mutually exhaustive set of units: e.g., states of the US. If we have a dataset containing all 50 states' values, it is not a random sample; it encompasses the entire population. We may want to allow for a state-specific intercept term, and the FE (a/k/a LSDV) estimator is a simple way to accomplish this.

Panel data methods for other data structures

We have considered the FD, FE and RE estimators as appropriate for strict panel data: those possessing both individual and time subscripts. But we may have datasets that do not possess a time element at all, but rather

a cross-sectional clustering variable (such as siblings within each family, or workers within each plant). Conceptually, we can apply any of these three panel data estimators in this context to take account of a common "family effect" or "plant effect". We cannot use `tsset` to declare such data as being panel data in Stata, but we can use the `i(panelvar)` option on any form of `xtreg` to designate the panel identifier. Just as the standard panel setup considers the likelihood that the individual's identity will give rise to unobserved heterogeneity in the form of a_i, we may consider it as likely that belonging to a particular family or working in a specific plant may have its own effect.

An alternative, available in most estimation commands in Stata, is the notion of *clustering*. We may consider families or plants in the prior example of *clusters*: groups within which

errors are likely to be correlated with one another. The *cluster covariance matrix* estimator allows for error variances to differ between clusters (but not within clusters), as well as allowing for correlations between errors in the same cluster (but not between clusters). Ignoring these correlations will cause estimated standard errors to be biased and inconsistent.

It may be invoked in `regress` and many other commands with the `,cluster(id)` option, where `id` specifies the name of an integer variable denoting cluster membership. The values of `id` need not be consecutive. When estimating cluster standard errors, it is important that there are more clusters than regressors in the model. In practical terms, this rules out the case that a panel identifier is specified as the cluster id and individual-specific constant terms are estimated. However, that does not rule out use of the cluster option in a FE mode because

that model does not literally estimate the N fixed effects among the regressors.

Seemingly unrelated regressions (SURE)

We often have a situation in which we want to estimate a similar specification for a number of different units: for instance, the estimation of a production function or cost function for each industry. If the equation to be estimated for a given unit meets the zero conditional mean assumption, we may estimate each equation independently. However, we may want to estimate the equations jointly: first, to allow cross-equation restrictions to be imposed or tested, and second, to gain efficiency, since we might expect the error terms across equations to be *contemporaneously correlated*.Such equations are often called *seemingly unrelated regressions*, and Zellner proposed an estimator for this problem: the SUR estimator. Unlike

the fixed effects and random effects estimators, whose large-sample justification is based on "small T, large N" datasets as $N \to \infty$, the SUR estimator is based on the large-sample properties of "large T, small N" datasets as $T \to \infty$. In that context, it may be considered a multiple time series estimator.

Equation i of the SUR model is:

$$y_i = X_i\beta_i + \epsilon_i, \;\; i = 1, \ldots, N \tag{8}$$

where y_i is the i^{th} equation's dependent variable and X_i is the matrix of regressors for the i^{th} equation, on which we have T observations. The disturbance process $\epsilon = [\epsilon_1', \epsilon_2', \ldots, \epsilon_N']'$ is assumed to have an expectation of zero and a covariance matrix of Ω. We will only consider the case where we have T observations per equation, although it is feasible to estimate the model with an unbalanced panel. Note also that although each X_i matrix will have T rows,

it may have K_i columns. Each equation may have a differing set of regressors, and apart from the constant term, there might be no variables in common across the X_i. Note that the application of SUR requires that the T observations per unit must exceed N, the number of units, in order to render Ω of full rank and invertible. If this constraint is not satisfied, SUR cannot be employed.

We assume that $E[\epsilon_{it}\epsilon_{js}] = \sigma_{ij},\ t = s$, otherwise zero. This implies that we are allowing for the error terms in different equations to be *contemporaneously correlated*, but assuming that they are not correlated at other points (including within a unit: they are assumed independent). Thus for any two error vectors,

$$\begin{aligned} E[\epsilon_i \epsilon_j'] &= \sigma_{ij} I_T \\ \Omega &= \Sigma \otimes I_T \end{aligned} \tag{9}$$

where Σ is the VCE of the N error vectors and $\otimes$ is the Kronecker matrix product (For any

matrices $A_{K\times L}, B_{M\times N}$, $A \otimes B = C_{KM\times LN}$. To form the product matrix, each element of A scalar multiplies the entire matrix B).

The efficient estimator for this problem is generalized least squares (GLS), in which we may write y as the stacked set of y_i vectors, and X as the block-diagonal matrix of X_i. Since the GLS estimator is

$$b_{GLS} = [X'\Omega^{-1}X][X'\Omega^{-1}y] \tag{10}$$

and

$$\Omega^{-1} = \Sigma^{-1} \otimes I \tag{11}$$

We can write the (infeasible) GLS estimator as

$$b_{GLS} = [X'(\Sigma^{-1} \otimes I)X]^{-1}[X'(\Sigma^{-1} \otimes I)y] \tag{12}$$

which if expanded demonstrates that each block of the $X_i'X_j$ matrix is weighted by the scalar σ_{ij}^{-1}. The large-sample VCE of b_{GLS} is the first term of this expression.

When will this estimator provide a gain in efficiency over equation-by-equation OLS? First, if the σ_{ij}, $i \neq j$ are actually zero, there is no gain. Second, if the X_i matrices are *identical* across equations—not merely having the same variable names, but containing the same numerical values—then GLS is identical to equation-by-equation OLS, and there is no gain. Beyond these cases, the gain in efficiency depends on the magnitude of the cross-equation contemporaneous correlations of the residuals. The higher are those correlations, the greater the gain. Furthermore, if the X_i matrices' columns are highly correlated across equations, the gains will be smaller.

The feasible SUR estimator requires a consistent estimate of Σ, the $N \times N$ contemporaneous covariance matrix of the equations' disturbance processes. The representative element σ_{ij}, the contemporaneous correlation be-

tween ϵ_i, ϵ_j, may be estimated from equation-by-equation OLS residuals as

$$s_{ij} = \frac{e_i' e_j}{T} \tag{13}$$

assuming that each unit's equation is estimated from T observations. These estimates are then used to perform the "Zellner step", where the algebra of partitioned matrices will show that the Kronecker products may be rewritten as products of the blocks in the expression for b_{GLS}. The estimator may be iterated. The GLS estimates will produce a new set of residuals, which may be used in a second Zellner step, and so on. Iteration will make the GLS estimates equivalent to maximum likelihood estimates of the system.

The SUR estimator is available in Stata via the `sureg` command. It is a panel data estimator applicable to data in the *wide* format. If the

data are set up in the long format more commonly used with panel data, the `reshape` command may be used to place them in the "wide" format. It is an attractive estimator relative to pooled OLS, or even in comparison with fixed effects, in that SUR allows each unit to have its own coefficient vector. Not only the constant term differs from unit to unit, but each of the slope parameters differ as well across units, as does σ_ϵ^2, which is constrained to be equal across units in pooled OLS, fixed effects or random effects estimators.

Standard F-tests may be used to compare the unrestricted SUR results with those that may be generated in the presence of linear constraints, such as cross-equation restrictions (see `constraint`). Cross-equation constraints correspond to the restriction that a particular regressor's effect is the same for each panel unit. The `isure` option may be used to iterate the estimates, as described above.

SUR with identical regressors

The second case discussed above in which SUR will generate the same point and interval estimates—the case of numerically identical regressors—arises quite often in economic theory and financial theory. For instance, the demand for each good should depend on the set of prices and income, or the portfolio share of assets held in a given class should depend on the returns to each asset and on total wealth. In this case, there is no reason to use anything other than OLS in terms of efficiency. However, SUR estimation is often employed in this case, since it allows for tests of cross-equation constraints, or estimation with those constraints in place.

If we try to apply SUR to a system with (numerically) identical regressors, such as a *complete set* of cost share or portfolio share equations, the SUR estimator will fail because the

error covariance matrix is singular. This holds not only for the unobservable errors, but also for the least squares residuals. A bit of algebra will show that if there are adding-up constraints across equations—for instance, if the set of y_i variables are a complete set of portfolio shares or demand shares—then the OLS residuals will sum to zero across equations, and their empirical covariance matrix will be singular *by construction*.

We may still want to utilize systems estimation in order to impose the cross-equation constraints arising from economic theory. In this case, the appropriate estimation strategy is to drop one of the equations and estimate the system of $(N-1)$ equations with SUR. The parameters of the N^{th} equation, in point and interval form, can be algebraically derived from those estimates. The feasible GLS estimates will be sensitive to which equation is dropped,

but iterated SUR will restore the invariance property of the maximum likelihood estimator of the problem.

Dynamic panel data models

A serious difficulty arises with the one-way fixed effects model in the context of a *dynamic panel data* (DPD) model: one containing a lagged dependent variable (and possibly other regressors), particularly in the "small T, large N" context. As Nickell (1981) shows, this arises because the demeaning process which subtracts the individual's mean value of y and each X from the respective variable creates a correlation between regressor and error. The mean of the lagged dependent variable contains observations 0 through $(T-1)$ on y, and the mean error—which is being conceptually subtracted from each ϵ_{it}—contains contemporaneous values of ϵ for $t = 1 \ldots T$. The resulting correlation creates a bias in the estimate of the coefficient of the lagged dependent variable which is

not mitigated by increasing N, the number of individual units. In the simplest setup of a pure $AR(1)$ model without additional regressors:

$$\begin{aligned} y_{it} &= \beta + \rho y_{i,t-1} + u_i + \epsilon_{it} \qquad (14) \\ y_{it} - y_{i\cdot} &= \rho(y_{i,t-1} - y_{i,\cdot-1}) + (\epsilon_{it} - \epsilon_{i\cdot}) \end{aligned}$$

The demeaning operation creates a regressor which *cannot* be distributed independently of the error term. Nickell demonstrates that the inconsistency of $\widehat{\rho}$ as $N \to \infty$ is of order $1/T$, which may be quite sizable in a "small T" context. If $\rho > 0$, the bias is invariably negative, so that the persistence of y will be underestimated. For reasonably large values of T, the limit of $(\widehat{\rho} - \rho)$ as $N \to \infty$ will be approximately $-(1+\rho)/(T-1)$: a sizable value, even if $T = 10$. With $\rho = 0.5$, the bias will be -0.167, or about 1/3 of the true value. The inclusion of additional regressors does not remove this bias. Indeed, if the regressors are correlated with the lagged dependent variable

to some degree, their coefficients may be seriously biased as well. Note also that this bias is not caused by an autocorrelated error process ϵ. The bias arises even if the error process is $i.i.d.$ If the error process is autocorrelated, the problem is even more severe given the difficulty of deriving a consistent estimate of the AR parameters in that context. The same problem affects the one-way random effects model. The u_i error component enters every value of y_{it} by assumption, so that the lagged dependent variable *cannot* be independent of the composite error process.

A solution to this problem involves taking first differences of the original model. Consider a model containing a lagged dependent variable and a single regressor X:

$$y_{it} = \beta_1 + \rho y_{i,t-1} + X_{it}\beta_2 + \ u_i + \epsilon_{it} \qquad (15)$$

The first difference transformation removes both the constant term and the individual effect:

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta X_{it}\beta_2 + \Delta \epsilon_{it} \qquad (16)$$

There is still correlation between the differenced lagged dependent variable and the disturbance process (which is now a first-order moving average process, or $MA(1)$): the former contains $y_{i,t-1}$ and the latter contains $\epsilon_{i,t-1}$. But with the individual fixed effects swept out, a straightforward instrumental variables estimator is available. We may construct instruments for the lagged dependent variable from the second and third lags of y, either in the form of differences or lagged levels. If ϵ is $i.i.d.$, those lags of y will be highly correlated with the lagged dependent variable (and its difference) but uncorrelated with the composite error process. Even if we had reason to believe that ϵ might be following an $AR(1)$ process, we could still follow this strategy, "backing off" one period and using the third and fourth lags of y

(presuming that the timeseries for each unit is long enough to do so).

The DPD (Dynamic Panel Data) approach of Arellano and Bond (1991) is based on the notion that the instrumental variables approach noted above does not exploit all of the information available in the sample. By doing so in a GMM context, we may construct more efficient estimates of the dynamic panel data model. The Arellano–Bond estimator can be thought of as an extension of the Anderson–Hsiao estimator implemented by `xtivreg, fd`. Arellano and Bond argue that the Anderson–Hsiao estimator, while consistent, fails to take all of the potential orthogonality conditions into account. Consider the equations

$$\begin{aligned} y_{it} &= X_{it}\beta_1 + W_{it}\beta_2 + v_{it} \\ v_{it} &= u_i + \epsilon_{it} \end{aligned} \tag{17}$$

where X_{it} includes strictly exogenous regressors, W_{it} are predetermined regressors (which

may include lags of y) and endogenous regressors, all of which may be correlated with u_i, the unobserved individual effect. First-differencing the equation removes the u_i and its associated omitted-variable bias. The Arellano–Bond estimator sets up a generalized method of moments (GMM) problem in which the model is specified as a system of equations, one per time period, where the instruments applicable to each equation differ (for instance, in later time periods, additional lagged values of the instruments are available). The instruments include suitable lags of the levels of the endogenous variables (which enter the equation in differenced form) as well as the strictly exogenous regressors and any others that may be specified. This estimator can easily generate an immense number of instruments, since by period τ all lags prior to, say, $(\tau - 2)$ might be individually considered as instruments. If T is nontrivial, it is often necessary to employ the

option which limits the maximum lag of an instrument to prevent the number of instruments from becoming too large. This estimator is available in Stata as `xtabond`.

A potential weakness in the Arellano–Bond DPD estimator was revealed in later work by Arellano and Bover (1995) and Blundell and Bond (1995). The lagged levels are often rather poor instruments for first differenced variables, especially if the variables are close to a random walk. Their modification of the estimator includes lagged levels as well as lagged differences. The original estimator is often entitled *difference GMM*, while the expanded estimator is commonly termed *System GMM*. The cost of the System GMM estimator involves a set of additional restrictions on the initial conditions of the process generating y.

Both the difference GMM and System GMM estimators have one-step and two-step variants. The two-step estimates of the difference GMM standard errors have been shown to have a severe downward bias. If the precision of the two-step estimators is to be evaluated for hypothesis tests, we should ensure that the "Windmeijer finite-sample correction" (see Windmeijer (2005)). to these standard errors has been applied. All of the features described above are available in David Roodman's improved version of official Stata's estimator. His version, `xtabond2`, offers a much more flexible syntax than official Stata's `xtabond`, which does not allow the same specification of instrument sets, nor does it provide the System GMM approach or the Windmeijer correction to the standard errors of the two-step estimates.