EC771: Econometrics, Spring 2010

Greene, Econometric Analysis (6th ed, 2008)

Chapters 14, 13: Systems of equations and panel data estimators

The seemingly unrelated regression estimator

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 for each industry. If the equation to be estimated for a given unit meets the conditions of being a proper OLS regression, we may of course estimate each equation independently. However, we may want to estimate the equations jointly: first, to allow tests of cross–equation restrictions to be

implemented, 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 (1962) proposed an estimator for this problem: the SURE.

We write the SUR model as

$$y_i = X_i\beta_i + \epsilon_i, \;\; i = 1, \ldots, M$$

where y_i is the i^{th} equation's dependent variable, on which we have T observations. We must stress that this need not be time–series data, although we often utilize the SUR model in that context. The error process $\epsilon = [\epsilon_1', \epsilon_2', \ldots, \epsilon_M']'$ 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 sample. 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, there may be no variables in common across the X_i. 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, $E[\epsilon_i \epsilon_j'] = \sigma_{ij} I_T$, and $\Omega = \Sigma \otimes I_T$.

The efficient estimator for this problem is generalized least squares, 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

$$\hat{\beta} = [\mathbf{X'}\Omega^{-1}\mathbf{X}][\mathbf{X'}\Omega^{-1}\mathbf{y}]$$

and

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

We can write the GLS estimator as

$$\hat{\beta} = [\mathbf{X'}(\Sigma^{-1} \otimes I)\mathbf{X}]^{-1}[\mathbf{X'}(\Sigma^{-1} \otimes I)\mathbf{y}]$$

which if expanded demonstrates that each block of the $X_i'X_j$ matrix is weighted by the scalar σ_{ij}^{-1}. The asymptotic covariance matrix of $\hat{\beta}$ is the first term of this expression.

When will this estimator provide a gain in efficiency? 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 OLS, and there is no gain. Beyond these cases, the gain in efficiency depends on the magnitude of the cross–equation 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 second case above—the case of 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 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.

The feasible SUR estimator requires a consistent estimate of σ_{ij}, which may be generated from OLS residuals via

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

assuming that each unit's equation is estimated from T observations. One could use a degrees

of freedom correction in the denominator, but relying on asymptotic properties, it is not necessary. These estimates are then used to perform the "Zellner step"; the algebra of partitioned matrices will show that the Kronecker products may be rewritten as products of the blocks in the expression for $\widehat{\beta}$. The estimator may be iterated: that is, 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.

Note that the application of SUR requires that the T observations per unit must exceed M, the number of units, in order to render Σ invertible. If this constraint is not satisfied, SUR cannot be employed, so that SUR is not an appropriate estimation technique for the "small T, large N" data structure that is often employed in, e.g., firm–level studies.

How might we test whether application of SUR has made a significant difference? A test for the diagonality of Σ was proposed by Breusch and Pagan (1980) (not to be confused with their test for heteroskedasticity): a Lagrange multiplier statistic which sums the squared correlations between residual vectors i and j, with a null hypothesis of diagonality (zero contemporaneous covariance between the errors of different equations).

The SUR estimator is available in Stata from the `sureg` command.

One special case should be noted: if we apply systems estimation to a system with (numerically) identical regressors, such as a 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 as well across equations, and their empirical covariance matrix will be singular by construction. We may still want to utilise 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 $M - 1$ equations with SUR. The parameters of the M^{th} equation can be 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.

SUR as a panel data estimator

The SUR estimator may be applied to panel, or longitudinal data, where we have T observations on each of M units. It is an attractive estimator, when it may be applied, in this context, since it automatically allows each unit to have its own coefficient vector—indeed, its own set of regressors, which may even differ in size from those of another unit—and its own error variance. 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. However, the estimator is not applicable to the "small T, large N" context in which we have a large number of units and a relatively short time series available for each unit. If the data are set up for SUR estimation, e.g. with each y_i as a separate variable, SUR may be applied directly. If they are not

in this "wide" format but rather in the "long" format, where a single y variable stacks each unit's y_i, then the Stata command `reshape` may be used to transform the data into the "wide" format necessary for SUR.

Pooled OLS

Estimators applicable to pooled cross–section / time–series data, or the special case of panel (longitudinal) data, are variations on the general model that might be applied to that sort of data structure:

$$y_{it} = \sum_{k=1}^{K} X_{kit}\beta_{kit} + \epsilon_{it},\ i = 1 \dots n,\ t = 1 \dots T.$$

A general and quite infeasible model, as it specifies $K \times n \times T$ regression coefficients to be estimated from only $n \times T$ observations (assuming that the XC/TS or panel data set is "balanced". Any feasible estimator must restrict the number of unknown coefficients to

far less than $n \times T$. One solution is posed by SUR, in which each cross–sectional unit has its own β vector; that removes the time–series dimension from the coefficient vector, leaving $K \times n$ coefficients. For SUR to be estimable, as we have seen, $n < T$. Another feasible estimator is "pooled OLS", in which the XC/TS nature of the data is merely ignored, and OLS applied to the data set; only K coefficients need be estimated. What might be wrong with pooled OLS? If cross–sectional datasets commonly contain heteroskedasticity, and time–series datasets may exhibit autocorrelation, it should be clear that a XC/TS dataset may contain both; and even its diagnosis is a nontrivial task in this setting, let alone its correction. Given the XC/TS nature of the data, robust estimators of the covariance matrix must be aware of the time-series dimension.

Fixed effects models

Let us rewrite the problem above to separate out the individual effect as

$$y_{it} = x'_{it}\beta + z'_i\alpha + \epsilon_{it}.$$

In this context, the x' does not contain a constant term; rather, the heterogeneity or individual effect is captured by z', which contains a constant term and possibly a number of other individual–specific factors. If z' contains only a constant term, then pooled OLS is a consistent and efficient estimator of $[\beta\ \alpha]$. However, it will often be the case that there are additional factors specific to the individual unit that must be taken into account; removing those variables from z' will cause the equation to be misspecified.

One of the most commonly applied estimators in this context is the "fixed effects" model,

which relaxes the assumption that the regression function is constant over time and space in a very modest way. A one–way fixed effects model permits each cross–sectional unit to have its own constant term. Unlike OLS, the slope estimates are constrained across units, as is the σ^2. This estimator is often termed the LSDV (least–squares dummy variable) model, since it is equivalent to including n dummy variables (or $n-1$, and retaining a conventional constant term) in the OLS regression. That model might be infeasible in many instances, where n could be in the thousands; an algebraic transformation may be applied to the data to work around that constraint of a very large regressor matrix. The (one–way) fixed effect model may then be written as

$$y_{it} = x'_{it}\beta + \alpha_i + \epsilon_{it}.$$

Alternatively, we might consider that the individual effect may be assumed to be indepen-

dent of x_i', and viewed as a draw from a distribution:

$$y_{it} = x_{it}'\beta + [u_i + \epsilon_{it}]$$

where the bracketed expression is a composite error term, with the u_i being a single draw per unit. This model could be consistently estimated by OLS, but that would be inefficient in not taking the nature of the composite disturbance process into account. To do so appropriately, we will develop the random effects estimator. This estimator will be efficient under the assumption that $X \perp u$, and inconsistent otherwise; thus a Hausman test may be used to compare the two alternatives of the one–way fixed effects estimator, which estimates a separate coefficient (constant term) for each unit, and the random effects estimator. The former will be consistent irregardless of the joint distribution of $[X\ u]$, whereas the

latter will only be consistent if they are independently distributed.

Let us first consider the one–way fixed effects model, and its extensions. This model may be written in matrix form as

$$y = X\beta + D\alpha + \epsilon$$

where D is a $nT \times n$ matrix of dummy variables d_i. The model has $K + n$ parameters (recalling that the β coefficients are all slopes) and when this number is too large to permit estimation, we rewrite the least squares solution as

$$b = (X'M_D X)^{-1}(X'M_D y)$$

where

$$M_D = I - D(D'D)^{-1}D'$$

is the idempotent matrix which is block–diagonal in $M_0 = I_T - T^{-1}\iota\iota'$. Premultiplying any data vector by M_0 performs the demeaning transformation: if we have a T–vector z_i, $M_0 z_i =$

$z_i - \bar{z}_i\iota$. The regression above estimates the slopes by the projection of demeaned y on demeaned X, without a constant term. The estimates a_i may be recovered from $a_i = \bar{y}_i - b'\bar{x}_i$, since for each unit, the regression surface passes through that unit's multivariate point of means. The asymptotic covariance matrix of b is $s^2[X'M_DX]^{-1}$, with s^2 based on the least squares residuals, but taking the proper degrees of freedom into account: $nT - n - K$.

When will this model have explanatory power? If and only if the variation of the individual's y above or below the individual's mean is significantly correlated with the variation of the individual's x values above or below the individual's vector of mean x values. For that reason, it is termed the "within estimator", since it depends on the variation within the unit. It does not matter if some individuals have, e.g., very high y values and very high X values, since it

is only the within variation that will show up as explanatory power. It follows that a variable must have variation within each individual to be usable in this format: e.g. in a dataset containing individuals' values, a variable such as gender or race cannot be used in a fixed–effects model, since it will have zero within variation for each individual, and its demeaned form will be a vector of zeros.

Since the model can be considered a regression containing a number of slopes and a complete set of dummy variables, an F–test may be used to examine the joint significance of those dummies: essentially, to test whether the α_i may be replaced by a single α coefficient, which represents $n-1$ restrictions on the model. This F test may be performed as a subset test, comparing the LSDV model with the pooled OLS model, and is routinely provided by software estimating the fixed effects model (as described below).

For data that have been `tsset`: that is, for which the individual ID variable and time–series calendar variable have been identified to Stata, as `tsset panelid timevar`—the one–way fixed effects model may be estimated via the command `xtreg yvar varlist, fe` where `fe` indicates that the (one–way) fixed–effects model is to be estimated. Unlike the SUR estimator, which generally requires a "balanced panel" (T observations on each of n units), the fixed effects estimator has no such restrictions; as long as there are at least two observations per unit, it may be applied. Since the individual fixed effect is in essence estimated from only those observations, the precision of that effect will depend on n_i. The command will print an estimate of the correlation between the unit–specific errors and the predicted values for the unit, which will figure in our later consideration of the random–effects estimator.

Another estimator that may be defined for this model is the "between estimator", in which the group means of y are regressed on the group means of X, in a regression of n observations. This estimator ignores all of the individual–specific variation in y and X that is considered by the within estimator, in effect replacing each observation for an individual by their mean behavior. This estimator has primarily pedagogical appeal, but may be readily applied to XC/TS data via `xtreg yvar varlist, be` as long as $n > K$. Obviously any variable that is constant over individuals (such as a time trend) cannot be included in the between estimator, since its average will not differ by individual.

One may show that the pooled OLS estimator is a matrix weighted average of the within and between estimators, with the weights defined by the relative precision of the two estimators.

One may ask, in a XC/TS context: where are the interesting sources of variation? In individuals' variation around their means, or in those means themselves? The within estimator takes account of (only) the former; the between estimator, (only) the latter.

The one–way fixed effects estimator may also be applied for time–specific effects rather than individual–specific effects. In Stata, this may readily be done by reversing the roles of the individual and time variables defined with `tsset`. Alternatively, one may apply a two–way fixed effects model, in which each individual and each time period has a constant term. Like the one–way model, this formulation is equivalent to including two sets of dummy variables in the equation (not more than one of which may be complete). If the number of time periods is relatively small, time dummy variables may be added to an `xtreg, fe` formulation. Stata does

not contain a command for two–way fixed effects, although that model is well defined: one essentially double–demeans the data:

$$y_{it}^{*} = y_{it} - \bar{y_i} - \bar{y_t} + \bar{\bar{y}}$$

where the last term is the grand mean of the series. The two–way FE model is equivalent to transforming the data in this fashion and projecting y^* on X^* without a constant term. The individual effects and time effects (and their standard errors) may be retrieved with some algebra. If one had a "small N, large T" dataset, a two–way FE model could be readily implemented by reversing the variables in `tsset` as described above and creating a set of individual dummies. This model treats the effect of being individual i at time t as the sum of those effects, which are assumed to be independent of one another. Nevertheless, this model is not frequently used, since we often may want to model the time variation more formally than with a simple dummy variable. Note

that in a model with time effects, one cannot have any variable that is constant over individuals, since demeaning over individuals will generate a zero vector. Thus, a variable such as a macro factor, which may be used in a individual FE model (since its mean value for each individual reflects time variation) cannot be included in a model with time effects. The time effect will capture all time variation that is not individual–specific.

Random effects models

Returning to the one–way (individual) fixed effects model, we allow for correlation between the fixed effect and regressors in that context. However, the resulting estimates should be considered as applying to the cross–sectional units under consideration, since they contain intercept terms labelled with those cross–sectional units' identifiers. In some cases, that may not

be problematic: e.g. if the units are the G-7 countries, we do not consider them a sample of major industrialized countries. In other circumstances, where the data may be survey data, that may be a difficulty. If the individual effects can be considered to be strictly independent of the regressors, then we might model the individual–specific constant terms (reflecting the unmodeled heterogeneity across units) to be draws from a distribution. This greatly reduces the number of parameters to be estimated, and conditional on that orthogonality, allows for inference to be made to the population from which the survey was constructed.

To implement the one–way random effects formulation

$$y_{it} = x'_{it}\beta + [u_i + \epsilon_{it}]$$

we assume that both ϵ and u are meanzero processes, distributed independent of X; that

they are each homoskedastic; that they are distributed independently of each other; and that each process represents independent realizations from its respective distribution, without correlation over individuals (nor time, for ϵ). For the T observations in the i^{th} block of data, we have the error process

$$\eta_{it} = u_i + \epsilon_{it}$$

which is the so–called error components model, with

$$E[\eta_{it}^2|X] = \sigma_\epsilon^2 + \sigma_u^2$$

and

$$E\eta_{it}\eta_{is}|X] = \sigma_u^2 \; t \neq s.$$

The covariance matrix of these T errors may then be seen to be

$$\Sigma = \sigma_\epsilon^2 I_T + \sigma_u^2 \iota_T \iota_T'.$$

Since observations i and j are independent, the full covariance matrix of η across the sample is block–diagonal in Σ: $\Omega = I_n \otimes \Sigma$.

The appropriate estimator for the slope parameters of this model is generalized least squares:

$$\widehat{\beta} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}y)$$
$$= \left(\sum_i X_i'\Sigma^{-1}X_i\right)^{-1}\left(\sum_i X_i'\Sigma^{-1}y_i\right)$$

(Note that this equation is in error in the text, p.295)

To compute this estimator, we require $\Omega^{-1/2} = [I_n \otimes \Sigma]^{-1/2}$, which involves

$$\Sigma^{-1/2} = \sigma_\epsilon^{-1}[I - T^{-1}\theta\iota_T\iota_T']$$

where

$$\theta = 1 - \frac{\sigma_\epsilon}{\sqrt{\sigma_\epsilon^2 + T\sigma_u^2}}$$

and the "quasi–demeaning" transformation defined by $\Sigma^{-1/2}$ is then $\sigma_\epsilon^{-1}(y_{it} - \theta\bar{y}_i)$: that is, rather than subtracting the entire individual mean of y, we should subtract some portion of it, as defined by θ. Compare this to the

LSDV model, in which we define the within estimator by setting $\theta = 1$. Like pooled OLS, the GLS random effects estimator is a matrix weighted average of the within and between estimators, but in this case applying optimal weights, as based on

$$\lambda = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + T\sigma_u^2} = (1 - \theta)^2$$

where λ is the weight attached to the covariance matrix of the between estimator. To the extent that λ differs from unity, pooled OLS will be inefficient, as it will attach too much weight on the between–units variation, attributing it all to the variation in X rather than apportioning some of the variation to the differences in u_i across units.

The setting $\lambda = 1$ ($\theta = 0$) is appropriate if $\sigma_u^2 = 0$, that is, if there are no random effects; then a pooled OLS model will be appropriate.

If $\lambda = 0$ ($\theta = 1$), then the appropriate estimator is the LSDV model of individual fixed effects.

To the extent that λ differs from zero, the within (LSDV) estimator will be inefficient, in that it applies zero weight to the between estimator. The random effects (GLS) estimator applies the optimal λ in the unit interval to the between estimator, whereas the fixed effects estimator imposes $\lambda = 0$. This would be appropriate if the variation in ϵ was trivial in comparison with the variation in u; then the dummy variables that identify each unit would, taken together, explain all of the variation in the composite error term.

How might we implement the feasible GLS estimator of the model? Since the fixed effects model is consistent, it may be used to generate OLS residuals, and an estimate of σ_ϵ^2.

Likewise, the pooled OLS model may generate a consistent estimate of $\sigma_\epsilon^2 + \sigma_u^2$. These two estimators may be used to define θ, and transform the data for the GLS model. (A possible complication: as generally defined, the two estimators above are not guaranteed to generate a positive estimate of σ_u^2 in finite samples. In that case, the variance estimates without degrees of freedom corrections, which will still be consistent, may be used. One might also question the orthogonality assumption in this case). A further complication: the GLS model, since it follows a quasi–demeaning approach, is capable of including variables that do not vary at the individual level (such as gender or race). But such variables cannot be included in the LSDV model, so that an alternative estimator must be defined, based on the between estimator's consistent estimate of $\sigma_u^2 + T^{-1}\sigma_\epsilon^2$. The same difficulty of a negative estimate of σ_u^2 may apply.

The feasible GLS estimator may be executed in Stata using the command `xtreg yvar varlist, re`. The command will display estimates of σ_u^2 and σ_ϵ^2, and what Stata calls ρ: the fraction of variance due to u_i. Breusch and Pagan have developed a Lagrange multiplier test for $\sigma_u^2 = 0$, which may be computed following a random–effects estimation via the command `xttest0`.

One may also estimate the parameters of the random–effects model with full maximum likelihood, and the `mle` option on the `xtreg, re` command requests that estimator. The application of MLE continues to assume that $X \perp u$, as well as assuming that the distributions of u and ϵ are Normal. This estimator will produce a likelihood ratio test of $\sigma_u^2 = 0$, corresponding to the Breusch–Pagan test available for the GLS estimator.

Testing for appropriateness of random effects

The Hausman specification test may be applied to evaluate the appropriateness of the assumption of orthogonality. If it is possible to apply the RE estimator, one would prefer it, since it is less costly in terms of degrees of freedom, does not require the omission of variables constant at the individual level, and does not require an assumption of $T \to \infty$ for consistency. On the other hand, the RE model will be inconsistent in the presence of correlation between X and u, and such a correlation—the unobserved heterogeneity depending on observed factors for the unit, such as some measure of scale—is quite plausible in practice. To implement the Hausman test, one estimates each form of the model, using the commands `estimates store set` after each estimation, with `set` defining that set of estimates: for instance, set might be "fe" for the fixed–effects model.

Then the command `hausman setconsist seteff` will invoke the Hausman test, where `setconsist` refers to the name of the fixed–effects estimates (which are consistent under the null and alternative) and `seteff` referring to the name of the random–effects estimates, which are only efficient under the null hypothesis of orthogonality. This test is based on the difference of the two estimated covariance matrices (which is not guaranteed to be positive definite) and the difference between the FE and RE slope coefficients' vectors. If the test cannot be computed, it casts doubt on the underlying specification of the model (not merely the issue of how the individual effects should be treated).

If the Hausman test indicates that the random effects cannot be considered orthogonal to the error, two approaches based on instrumental variables estimation may be utilized.

One, the Hausman–Taylor estimator, assumes that some of the regressors in X are correlated with u, but that none are correlated with ϵ. This estimator is available in Stata as `xthtaylor`. Their approach is based on the notion that we can divide the regressors into four categories: the interaction of time varying (x) / time invariant (z) and uncorrelated with u_i (1) / correlated (2). For example, x_2 are those time-varying regressors that are thought to be correlated with u_i. The time–invariant variables have a coefficient vector α, while the time–varying regressors have the coefficient vector β. The demeaning (fixed effects) transformation may be used on the time varying variables, even for those correlated with u_i, to estimate the β subset of the coefficient vector. Its residual variance estimator is consistent for σ_ϵ^2. The group means of the residuals from the fixed–effects regression are used as the dependent variable in a instrumental variables regression

on the z variables with instrumental variables z_1 and x_1 (the time–invariant variables are repeated T times). The identification requirement forces K_1 (the number of x_1 variables) to be at least as large as L_2 (the number of z_2 variables). This regression provides a consistent estimate of α, and its residual variance is a consistent estimate of the variance of the composite error term, giving the weight θ for feasible GLS. The application of the H–T estimator circumvents the problem of x_2 and z_2 variables being potentially correlated with u_i, but requires that we can identify variables of type 1 that are surely not correlated with the random effects.

There is also a IV form of the random effects estimator, `xtivreg`, which permits correlation between some of the X variables and the idiosyncratic error ϵ. These are quite different assumptions about the nature of any suspected correlation between regressor and the

composite error term, and should be selected with caution. The `xtivreg` command also supports fixed–effects, between–effects, and first–differenced estimators in an instrumental variables context.

Dynamic panel data models

We have discussed the concern that a random–effects formulation may not be appropriate due to correlation of regressor and the random effect: a problem mitigated by the development of the Hausman–Taylor approach. One always may eschew the random–effects formulation and utilise the fixed–effects formulation in its place, and in static regression models, the latter will always yield consistent estimates. However, a serious difficulty arises with the one–way fixed effects model in the context of a dynamic panel data model: one containing a lagged dependent variable (and possibly other regressors), particularly in the "small

T, large N" context. As Nickell (Econometrica, 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, as Nickell shows, creates a bias in the estimate of the coefficient of the lagged dependent variable which is not affected by the number of individuals n. In the simplest setup, of a pure AR model without additional regressors:

$$y_{it} = \beta + \rho y_{i,t-1} + f_i + \epsilon_{it}$$
$$y_{it} - y_{i\cdot} = \rho(y_{i,t-1} - y_{i,\cdot-1}) + (\epsilon_{it} - \epsilon_{i\cdot})$$

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 plim $(\widehat{\rho} - \rho)$ will be approximately $\frac{-(1+\rho)}{T-1}$: 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: even if the error process is i.i.d., the bias arises. 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: since the u_i error component enters every value of y_{it} by assumption, the lagged dependent variable cannot be independent of the composite error process.

A solution to this problem (not without its own complications) involves taking first differences of the original model. Consider a model containing a lagged dependent variable and a single regressor x:

$$y_{it} = \beta_0 + \rho y_{i,t-1} + x_{it}\beta_1 + \ f_i + \epsilon_{it}$$

Applying the first difference transformation yields:

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta x_{it}\beta + \Delta \epsilon_{it}$$

There is still correlation between the differenced LDV and the disturbance process (which is now a MA(1) error): 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 LDV 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 LDV (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 approach of Arellano and Bond (R.E. Studies, 1991), Arellano and Bover (J. Econometrics, 1995) and Blundell and Bond (J. Econometrics, 1998) is based on the notion that the instrumental variables approach noted above does not exploit all of the information available in the sample, and that by doing so, in

a GMM context, one may construct more efficient estimates of the dynamic panel data model. The A–B estimator can be thought of as an outgrowth of H–T, in which the x_2 variables (time–varying measures correlated with u_i) include y_{t-1}. A–B argue that the IV estimator does not take all of the potential orthogonality conditions into account. Consider the equations

$$y_{it} = x_{it}\beta_1 + w_{it}\beta_2 + u_{it}$$
$$u_{it} = v_i + \epsilon_{it}$$

where x_{it} are strictly exogenous covariates, w_{it} are predetermined covariates (which may include lags of y) and endogenous covariates, all of which may be correlated with v_i and v_i are unobserved individual effects. First–differencing the equation removes the v_i and its associated omitted–variable bias. The Arellano–Bond estimator sets up a 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, suitably 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 T all lags prior to, say, $T-2$ might be individually considered as instruments. If T is nontrivial, it is often necessary to exercise one of the options which limits the maximum lag of an instrument to prevent the number of instruments from becoming too large.

A potential weakness in the A–B estimator was revealed in later work by Arellano–Bover and Blundell–Bond: the lagged levels are often rather poor instruments for first differenced

variables, especially if the variables are close to a random walk. Thus, 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". Both 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. When the two–step estimators are to be reported, one should ensure that the "Windmeijer finite–sample correction" to these standard errors has been applied.

All of the features described above are available in 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 nor the essential Windmeijer correction.

An excellent guide to the DPD estimators is provided in Steve Bond's "Dynamic panel data models: a guide to microdata methods and practice", available from EconPapers (CeMMAP working paper 09/02 at Institute for Fiscal Studies).