Management and analysis of panel data in economics and finance

Christopher F Baum

Boston College and DIW Berlin

January 2009

Panel or longitudinal data are widely available in many fields of economics and finance. Econometric analysis using panel data can make use of estimators which can yield results more powerful than those available from pure cross-section or time-series data.

Panel or longitudinal data are widely available in many fields of economics and finance. Econometric analysis using panel data can make use of estimators which can yield results more powerful than those available from pure cross-section or time-series data.

However, with this power we face a number of challenges in handling the data and dealing with additional econometric issues—such as unobserved heterogeneity—that must be properly handled to provide consistent estimates.

Panel or longitudinal data are widely available in many fields of economics and finance. Econometric analysis using panel data can make use of estimators which can yield results more powerful than those available from pure cross-section or time-series data.

However, with this power we face a number of challenges in handling the data and dealing with additional econometric issues—such as unobserved heterogeneity—that must be properly handled to provide consistent estimates.

In this lecture, we will touch upon some of these issues, and discuss hands-on solutions for some of the data management issues that arise in the context of panel data.

The discussion that follows is presented in much greater detail in three sources:

- *An Introduction to Modern Econometrics Using Stata*, Baum, C.F., Stata Press, 2006 (particularly Chapter 8).

The discussion that follows is presented in much greater detail in three sources:

- *An Introduction to Modern Econometrics Using Stata*, Baum, C.F., Stata Press, 2006 (particularly Chapter 8).
- *An Introduction to Stata Programming*. Baum, C.F., Stata Press, 2009.

The discussion that follows is presented in much greater detail in three sources:

- *An Introduction to Modern Econometrics Using Stata*, Baum, C.F., Stata Press, 2006 (particularly Chapter 8).
- *An Introduction to Stata Programming*. Baum, C.F., Stata Press, 2009.
- How to do xtabond2. Roodman, D. Forthcoming, *Stata Journal*.
  `http://ideas.repec.org/p/boc/asug06/8.html`

Forms of panel data

To define the problems of data management, consider a dataset in which we have k variables each with T time-series observations. The second dimension of panel data need not be calendar time, but many estimation techniques assume that it can be treated as such, so that operations such as first differencing make sense.

Forms of panel data

To define the problems of data management, consider a dataset in which we have k variables each with T time-series observations. The second dimension of panel data need not be calendar time, but many estimation techniques assume that it can be treated as such, so that operations such as first differencing make sense.

These data may be commonly stored in either the *long form* or the *wide form*, in Stata parlance. In the long form, each observation has both an i and t subscript.

Long form data:

```
. list, noobs sepby(state)
```

state	year	pop
CT	1990	3291967
CT	1995	3324144
CT	2000	3411750
MA	1990	6022639
MA	1995	6141445
MA	2000	6362076
RI	1990	1005995
RI	1995	1017002
RI	2000	1050664

However, you often encounter data in the wide form, in which different variables (or columns of the data matrix) refer to different time periods.

However, you often encounter data in the wide form, in which different variables (or columns of the data matrix) refer to different time periods.

Wide form data:

```
. list, noobs
```

state	pop1990	pop1995	pop2000
CT	3291967	3324144	3411750
MA	6022639	6141445	6362076
RI	1005995	1017002	1050664

However, you often encounter data in the wide form, in which different variables (or columns of the data matrix) refer to different time periods.

Wide form data:

```
. list, noobs
```

state	pop1990	pop1995	pop2000
CT	3291967	3324144	3411750
MA	6022639	6141445	6362076
RI	1005995	1017002	1050664

In a variant on this theme, the wide form data could also index the observations by the time period, and have the same measurement for different units stored in different variables.

The former kind of wide-form data, where time periods are arrayed across the columns, is often found in spreadsheets or on-line data sources.

The former kind of wide-form data, where time periods are arrayed across the columns, is often found in spreadsheets or on-line data sources.

These examples illustrate a *balanced panel*, where each unit is represented in each time period. That is often not available, as different units may enter and leave the sample in different periods (companies may start operations or liquidate, household members may die, etc.) In those cases, we must deal with *unbalanced panels*. Stata's data transformation commands are uniquely handy in that context.

Data management for panel data

The data management challenge: for most purposes of data transformation, estimation and graphing, data are more easily used in the long form. Stata constructs such as the by-group require that we have the data stored that way. So if we have wide form data, how do we get there from here?

The solution to this problem is Stata's `reshape` command, an immensely powerful tool for reformulating a dataset in memory without recourse to external files. In statistical packages lacking a data-reshape feature, common practice entails writing the data to one or more external text files and reading it back in. With the proper use of `reshape`, this is not necessary in Stata. But `reshape` requires, first of all, that the data to be reshaped are labelled in such a way that they can be handled by the mechanical rules that the command applies. In situations beyond the simple application of `reshape`, it may require some experimentation to construct the appropriate command syntax. This is all the more reason for enshrining that code in a do-file as some day you are likely to come upon a similar application for `reshape`.

The `reshape` command works with the notion of $x_{i,j}$ data. Its syntax lists the variables to be stacked up, and specifies the *i* and *j* variables, where the *i* variable indexes the rows and the *j* variable indexes the columns in the existing form of the data. If we have a dataset in the wide form, with time periods incorporated in the variable names, we could use

```
. reshape long expp revpp avgsal math4score math7score, i(distid) j(year)
(note: j = 1992 1994 1996 1998)
Data                                wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                       550   ->    2200
Number of variables                   21   ->       7
j variable (4 values)                      ->   year
xij variables:
         expp1992 expp1994 ... expp1998    ->   expp
      revpp1992 revpp1994 ... revpp1998    ->   revpp
   avgsal1992 avgsal1994 ... avgsal1998    ->   avgsal
math4score1992 math4score1994 ... math4score1998->math4score
math7score1992 math7score1994 ... math7score1998->math7score
-----------------------------------------------------------------------------
```

You use `reshape long` because the data are in the wide form and we want to place them in the long form. You provide the variable names to be stacked *without* their common suffixes: in this case, the `year` embedded in their wide-form variable name. The `i` variable is `distid` and the `j` variable is `year`: together, those variables uniquely identify each measurement. Stata's description of `reshape` speaks of *i* defining a unique observation and *j* defining a subobservation logically related to that observation. Any additional variables that do not vary over *j* are not specified in the `reshape` statement, as they will be automatically replicated for each *j*.

What if you wanted to reverse the process, and translate the data from the long to the wide form?

What if you wanted to reverse the process, and translate the data from the long to the wide form?

```
. reshape wide expp revpp avgsal math4score math7score, i(distid) j(year)
(note: j = 1992 1994 1996 1998)
Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                     2200   ->     550
Number of variables                   7   ->      21
j variable (4 values)              year   ->   (dropped)
xij variables:
                                   expp   ->   expp1992 expp1994 ... expp1998
                                  revpp   ->   revpp1992 revpp1994 ... revpp199
> 8
                                 avgsal   ->   avgsal1992 avgsal1994 ... avgsal
> 1998
                             math4score   ->   math4score1992 math4score1994 ..
> . math4score1998
                             math7score   ->   math7score1992 math7score1994 ..
> . math7score1998
-----------------------------------------------------------------------------
```

This example highlights the importance of having appropriate variable names for `reshape`. If our wide-form dataset contained the variables `expp1992, Expen94, xpend_96` and `expstu1998` there would be no way to specify the common stub labeling the choices. However, one common case can be handled without the renaming of variables. Say that we have the variables `exp92pp, exp94pp, exp96pp, exp98pp`. The command

```
reshape long exp@pp, i(distid) j(year)
```

will deal with that case, with the `@` as a placeholder for the location of the *j* component of the variable name.

This discussion has only scratched the surface of `reshape`'s capabilities. There is no substitute for experimentation with this command after a careful perusal of `help reshape`, as it is one of the most complicated elements of Stata.

This discussion has only scratched the surface of `reshape`'s capabilities. There is no substitute for experimentation with this command after a careful perusal of `help reshape`, as it is one of the most complicated elements of Stata.

When working with panel data, we also must consider *combining* datasets, as often data are available for one panel at a time (for instance, cross-sectional information on the 100 largest companies at each year-end). In this next section, we take up that issue.

Combining datasets

You may be aware that Stata can only work with one dataset at a time. How, then, do you combine datasets in Stata? First of all, it is important to understand that at least one of the datasets to be combined must already have been saved in Stata format. Second, you should realize that each of Stata's commands for combining datasets provides a certain functionality, which should not confused with that of other commands.

Combining datasets

You may be aware that Stata can only work with one dataset at a time. How, then, do you combine datasets in Stata? First of all, it is important to understand that at least one of the datasets to be combined must already have been saved in Stata format. Second, you should realize that each of Stata's commands for combining datasets provides a certain functionality, which should not confused with that of other commands.

For instance, consider the `append` command with two stylized datasets:

$$\text{dataset1}: \begin{pmatrix} id & var1 & var2 \\ 112 & \vdots & \vdots \\ 216 & \vdots & \vdots \\ 449 & \vdots & \vdots \end{pmatrix}$$

$$\text{dataset2}: \begin{pmatrix} id & var1 & var2 \\ 126 & \vdots & \vdots \\ 309 & \vdots & \vdots \\ 421 & \vdots & \vdots \\ 604 & \vdots & \vdots \end{pmatrix}$$

These two datasets contain the same variables, as they must for `append` to sensibly combine them. If `dataset2` contained `idcode, Var1, Var2` the two datasets could not sensibly be appended without renaming the variables.[1] Appending these two datasets with common variable names creates a single dataset containing all of the observations:

[1] Recall that in Stata `var1` and `Var1` are two separate variables.

$$\text{combined} : \begin{pmatrix} id & var1 & var2 \\ 112 & \vdots & \vdots \\ 216 & \vdots & \vdots \\ 449 & \vdots & \vdots \\ 126 & \vdots & \vdots \\ 309 & \vdots & \vdots \\ 421 & \vdots & \vdots \\ 604 & \vdots & \vdots \end{pmatrix}$$

$$\text{combined}: \begin{pmatrix} id & var1 & var2 \\ 112 & \vdots & \vdots \\ 216 & \vdots & \vdots \\ 449 & \vdots & \vdots \\ 126 & \vdots & \vdots \\ 309 & \vdots & \vdots \\ 421 & \vdots & \vdots \\ 604 & \vdots & \vdots \end{pmatrix}$$

The rule for `append`, then, is that if datasets are to be combined, they should share the same variable names and datatypes (string vs. numeric). In the above example, if `var1` in `dataset1` was a `float` while that variable in `dataset2` was a `string` variable, they could not be appended.

It is permissible to append two datasets with differing variable names in the sense that `dataset2` could also contain an additional variable or variables (for example, `var3, var4`). The values of those variables in the observations coming from `dataset1` would then be set to missing.

It is permissible to append two datasets with differing variable names in the sense that `dataset2` could also contain an additional variable or variables (for example, `var3, var4`). The values of those variables in the observations coming from `dataset1` would then be set to missing.

While `append` combines datasets by adding observations to the existing variables, the other key command, `merge` combines variables for the existing observations.

Consider these two stylized datasets:

$$\text{dataset1}: \begin{pmatrix} id & var1 & var2 \\ 112 & \vdots & \vdots \\ 216 & \vdots & \vdots \\ 449 & \vdots & \vdots \end{pmatrix}$$

$$\text{dataset3}: \begin{pmatrix} id & var22 & var44 & var46 \\ 112 & \vdots & \vdots & \vdots \\ 216 & \vdots & \vdots & \vdots \\ 449 & \vdots & \vdots & \vdots \end{pmatrix}$$

We may `merge` these datasets on the common *merge key*: in this case, the `id` variable:

$$\text{combined}: \left(\begin{array}{cccccc} \textit{id} & \textit{var}1 & \textit{var}2 & \textit{var}22 & \textit{var}44 & \textit{var}46 \\ 112 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 216 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 449 & \vdots & \vdots & \vdots & \vdots & \vdots \end{array} \right)$$

The rule for `merge`, then, is that if datasets are to be combined on one or more *merge keys*, they each must have one or more variables with a common name and datatype (string vs. numeric). In the example above, each dataset must have a variable named `id`. That variable can be numeric or string, but that characteristic of the merge key variables must match across the datasets to be merged. Of course, we need not have exactly the same observations in each dataset: if `dataset3` contained observations with additional `id` values, those observations would be merged with missing values for `var1` and `var2`.

The rule for `merge`, then, is that if datasets are to be combined on one or more *merge keys*, they each must have one or more variables with a common name and datatype (string vs. numeric). In the example above, each dataset must have a variable named `id`. That variable can be numeric or string, but that characteristic of the merge key variables must match across the datasets to be merged. Of course, we need not have exactly the same observations in each dataset: if `dataset3` contained observations with additional `id` values, those observations would be merged with missing values for `var1` and `var2`.

This is the simplest kind of merge: the *one-to-one merge*. Stata supports several other types of merges. But the key concept should be clear: the `merge` command combines datasets "horizontally", adding variables' values to existing observations.

The long-form dataset is very useful if you want to add aggregate-level information to individual records. For instance, we may have panel data for a number of companies for several years. We may want to attach various macro indicators (interest rate, GDP growth rate, etc.) that vary by year but not by company. We would place those macro variables into a dataset, indexed by year, and sort it by year.

The long-form dataset is very useful if you want to add aggregate-level information to individual records. For instance, we may have panel data for a number of companies for several years. We may want to attach various macro indicators (interest rate, GDP growth rate, etc.) that vary by year but not by company. We would place those macro variables into a dataset, indexed by year, and sort it by year.

We could then `use` the firm-level panel dataset and sort it by `year`. A `merge` command can then add the appropriate macro variables to each instance of `year`. This use of `merge` is known as a *one-to-many* match merge, where the `year` variable is the *merge key*.

The long-form dataset is very useful if you want to add aggregate-level information to individual records. For instance, we may have panel data for a number of companies for several years. We may want to attach various macro indicators (interest rate, GDP growth rate, etc.) that vary by year but not by company. We would place those macro variables into a dataset, indexed by year, and sort it by year.

We could then `use` the firm-level panel dataset and sort it by `year`. A `merge` command can then add the appropriate macro variables to each instance of `year`. This use of `merge` is known as a *one-to-many* match merge, where the `year` variable is the *merge key*.

Note that the merge key may contain several variables: we might have information specific to industry and year that should be merged onto each firm's observations.

By default, `merge` creates a new variable `_merge`, which takes on integer values for each observation of 1 if that observation was only found in the master dataset, 2 if it was only found in the using dataset, or 3 if it was found in both datasets. In this case, we expect that `tab _merge` should reveal that all values equal 3. We can also use the `uniqusing` option to ensure that there are no duplicate values of `year` in the using file, as a duplicate value of `distid` must be a data entry error. If the same `year` mistakenly appears on two records in the using file, asserting `uniqusing` will cause `merge` to fail.

By default, `merge` creates a new variable `_merge`, which takes on integer values for each observation of 1 if that observation was only found in the master dataset, 2 if it was only found in the using dataset, or 3 if it was found in both datasets. In this case, we expect that `tab _merge` should reveal that all values equal 3. We can also use the `uniqusing` option to ensure that there are no duplicate values of `year` in the using file, as a duplicate value of `distid` must be a data entry error. If the same `year` mistakenly appears on two records in the using file, asserting `uniqusing` will cause `merge` to fail.

You may also use a `uniqmaster` option, where the master file should contain only one record for the merge key (which may include several variables), or the `unique` option in the case of the one-to-one merge where there should be a perfect match between the two files.

In your particular application, you may find that `_merge` values of 1 or 2 are appropriate. The key notion is that you should always tabulate `_merge` and consider whether the results of the merge are sensible in the context of your work. It is an excellent idea to use the `uniqmaster`, `uniqusing` or `unique` options on the `merge` command whenever those conditions should logically be satisfied in your data.

In your particular application, you may find that `_merge` values of 1 or 2 are appropriate. The key notion is that you should always tabulate `_merge` and consider whether the results of the merge are sensible in the context of your work. It is an excellent idea to use the `uniqmaster`, `uniqusing` or `unique` options on the `merge` command whenever those conditions should logically be satisfied in your data.

In comparison with a lengthy and complicated do-file using a set of `replace` statements, the `merge` technique is far better. This technique proves exceedingly useful when working with individual data and panel data where we have aggregate information to be combined with the individual-level data.

There are very good reasons to employ a one-to-many merge, as we did above with macro variables, or its inverse: a many-to-one merge, which would essentially reverse the roles of the master and using datasets. But there is a great danger in stumbling into the alternative to the one-to-many or one-to-one merge: the *many-to-many* merge. This problem arises when there are multiple observations in both datasets for some values of the merge key variable(s).

The result of match-merging two datasets which both have more than one value of the merge key variable(s) is unpredictable, as it depends on the sort order of the datasets. This leads to the seemingly illogical result that repeated execution of the same `do`–file will most likely result in a different number of cases in the result dataset without any error indication. There is no unique outcome for a many-to-many merge. When it is encountered it usually results from a coding error in one of the files.

The result of match-merging two datasets which both have more than one value of the merge key variable(s) is unpredictable, as it depends on the sort order of the datasets. This leads to the seemingly illogical result that repeated execution of the same `do`–file will most likely result in a different number of cases in the result dataset without any error indication. There is no unique outcome for a many-to-many merge. When it is encountered it usually results from a coding error in one of the files.

Stata's `duplicates` command is very useful in tracking down such errors. To prevent such difficulties in employing `merge`, you should specify either the `uniqmaster` or the `uniqusing` option in a match merge. If no `uniq`...option is used, observations may be matched inappropriately.

Estimation for panel data

We first consider estimation of models that satisfy the zero conditional mean assumption for OLS regression: that is, the conditional mean of the error process, conditioned on the regressors, is zero. This does not rule out non-$i.i.d$. errors, but it does rule out endogeneity of the regressors and, generally, the presence of lagged dependent variables. We will deal with these exceptions later.

Estimation for panel data

We first consider estimation of models that satisfy the zero conditional mean assumption for OLS regression: that is, the conditional mean of the error process, conditioned on the regressors, is zero. This does not rule out non-$i.i.d$. errors, but it does rule out endogeneity of the regressors and, generally, the presence of lagged dependent variables. We will deal with these exceptions later.

The most commonly employed model for panel data, the *fixed effects* estimator, addresses the issue that no matter how many individual-specific factors you may include in the regressor list, there may be *unobserved heterogeneity* in a pooled OLS model. This will generally cause OLS estimates to be biased and inconsistent.

Given longitudinal data $\{y\ X\}$, each element of which has two subscripts: the unit identifier i and the time identifier t, we may define a number of models that arise from the most general linear representation:

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

Given longitudinal data $\{y\ X\}$, each element of which has two subscripts: the unit identifier i and the time identifier t, we may define a number of models that arise from the most general linear representation:

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

Assume a balanced panel of $N \times T$ observations. Since this model contains $K \times N \times T$ regression coefficients, it cannot be estimated from the data. We could ignore the nature of the panel data and apply pooled ordinary least squares,pooled OLS which would assume that $\beta_{kit} = \beta_k\ \forall\ k, i, t$, but that model might be viewed as overly restrictive and is likely to have a very complicated error process (e.g., heteroskedasticity across panel units, serial correlation within panel units, and so forth). Thus the pooled OLS solution is not often considered to be practical.

One set of panel data estimators allow for heterogeneity across panel units (and possibly across time), but confine that heterogeneity to the intercept terms of the relationship. These techniques, the *fixed effects* and *random effects* models, we consider below. They impose restrictions on the model above of $\beta_{kit} = \beta_k \; \forall i, t, \; k > 1$, assuming that β_1 refers to the constant term in the relationship.

An alternative technique which may be applied to "small N, large T" panels is the method of *seemingly unrelated regressions* or SURE. The "small N, large T" setting refers to the notion that we have a relatively small number of panel units, each with a lengthy time series: for instance, financial variables of the ten largest U.S. manufacturing firms, observed over the last 40 calendar quarters. The SURE technique (implemented in Stata as `sureg`) requires that the number of time periods exceeds the number of cross-sectional units.

The general structure above may be restricted to allow for heterogeneity across units without the full generality (and infeasibility) that this equation implies. In particular, we might restrict the slope coefficients to be constant over both units and time, and allow for an intercept coefficient that varies by unit or by time. For a given observation, an intercept varying over units results in the structure:

$$y_{it} = \sum_{k=2}^{K} X_{kit}\beta_k + u_i + \epsilon_{it} \tag{2}$$

There are two interpretations of u_i in this context: as a parameter to be estimated in the model (a so-called *fixed effect*) or alternatively, as a component of the disturbance process, giving rise to a composite error term $[u_i + \epsilon_{it}]$: a so-called *random effect*. Under either interpretation, u_i is taken as a random variable.

There are two interpretations of u_i in this context: as a parameter to be estimated in the model (a so-called *fixed effect*) or alternatively, as a component of the disturbance process, giving rise to a composite error term $[u_i + \epsilon_{it}]$: a so-called *random effect*. Under either interpretation, u_i is taken as a random variable.

If we treat it as a fixed effect, we assume that the u_i may be correlated with some of the regressors in the model. The fixed-effects estimator removes the fixed-effects parameters from the estimator to cope with this incidental parameter problem, which implies that all inference is conditional on the fixed effects in the sample. Use of the random effects model implies additional orthogonality conditions—that the u_i are not correlated with the regressors—and yields inference about the underlying population that is not conditional on the fixed effects in our sample.

We could treat a time-varying intercept term similarly: as either a fixed effect (giving rise to an additional coefficient) or as a component of a composite error term. We concentrate here on so-called *one-way fixed (random) effects* models in which only the individual effect is considered in the "large N, small T" context most commonly found in economic and financial research. Stata's set of `xt` commands include those which extend these panel data models in a variety of ways. For more information, see `help xt`.

One-way fixed effects: the within estimator

Rewrite the equation to express the individual effect u_i as

$$y_{it} = X^*_{it}\beta^* + Z_i\alpha + \epsilon_{it} \tag{3}$$

In this context, the X^* matrix does not contain a units vector. The heterogeneity or individual effect is captured by Z, which contains a constant term and possibly a number of other individual-specific factors. Likewise, β^* contains $\beta_2 \ldots \beta_K$ from the equation above, constrained to be equal over i and t. If Z contains only a units vector, 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, and omitting those variables from Z will cause the equation to be misspecified.

The *fixed effects* model deals with this problem by relaxing 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 while the slope estimates (β^*) are constrained across units, as is the σ_ϵ^2. This estimator is often termed the *LSDV* (least-squares dummy variable) model, since it is equivalent to including ($N - 1$) dummy variables in the OLS regression of y on X (including a units vector). The *LSDV* model may be written in matrix form as:

$$y = X\beta + D\alpha + \epsilon \tag{4}$$

where D is a $NT \times N$ matrix of dummy variables d_i (assuming a balanced panel of $N \times T$ observations).

The model has $(K-1)+N$ 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

$$b = (X'M_D X)^{-1}(X'M_D y) \tag{5}$$

where

$$M_D = I - D(D'D)^{-1}D' \tag{6}$$

is an idempotent matrix which is block–diagonal in $M_0 = I_T - T^{-1}\iota\iota'$ (ι a T–element units vector).

The model has $(K-1)+N$ 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

$$b = (X'M_D X)^{-1}(X'M_D y) \tag{5}$$

where

$$M_D = I - D(D'D)^{-1}D' \tag{6}$$

is an idempotent matrix which is block–diagonal in $M_0 = I_T - T^{-1}\iota\iota'$ (ι a T–element units vector).

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. This is a generalization of the OLS result that in a model with a constant term the regression surface passes through the *entire sample's* multivariate point of means.

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. This is a generalization of the OLS result that in a model with a constant term the regression surface passes through the *entire sample's* multivariate point of means.

The large-sample VCE 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: $NT - N - (K - 1)$.

This model will 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.

This model will 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. This is the panel analogue to the notion that OLS on a cross-section does not seek to "explain" the mean of y, but only the variation around that mean.

This has the clear implication that any characteristic which does not vary over time for each *unit* cannot be included in the model: for instance, an individual's gender, or a firm's three-digit SIC (industry) code. The unit-specific intercept term absorbs all heterogeneity in y and X that is a function of the identity of the unit, and any variable constant over time for each unit will be perfectly collinear with the unit's indicator variable.

The one-way individual fixed effects model may be estimated by the Stata command [XT] **xtreg** using the `fe` (fixed effects) option. The command has a syntax similar to `regress`:

`xtreg` *depvar indepvars*`,` `fe` [*options*]

The one-way individual fixed effects model may be estimated by the Stata command [XT] **xtreg** using the `fe` (fixed effects) option. The command has a syntax similar to `regress`:

```
xtreg depvar indepvars, fe [options]
```

As with standard regression, options include `robust` and `cluster()`. The command output displays estimates of σ_u^2 (labeled `sigma_u`), σ_ϵ^2 (labeled `sigma_e`), and what Stata terms `rho`: the fraction of variance due to u_i. Stata estimates a model in which the u_i of Equation (2) are taken as deviations from a single constant term, displayed as `_cons`; therefore testing that all u_i are zero is equivalent in our notation to testing that all α_i are identical. The empirical correlation between u_i and the regressors in X^* is also displayed as `corr(u_i, Xb)`.

The fixed effects estimator does not require a balanced panel. As long as there are at least two observations per unit, it may be applied. However, since the individual fixed effect is in essence estimated from the observations of each unit, the precision of that effect (and the resulting slope estimates) will depend on N_i.

The fixed effects estimator does not require a balanced panel. As long as there are at least two observations per unit, it may be applied. However, since the individual fixed effect is in essence estimated from the observations of each unit, the precision of that effect (and the resulting slope estimates) will depend on N_i.

We wish to test whether the individual-specific heterogeneity of α_i is necessary: are there distinguishable intercept terms across units? `xtreg,fe` provides an F-test of the null hypothesis that the constant terms are equal across units. If this null is rejected, pooled OLS would represent a misspecified model. The one-way fixed effects model also assumes that the errors are not contemporaneously correlated across units of the panel. This hypothesis can be tested (provided $T > N$) by the Lagrange multiplier test of Breusch and Pagan, available as the author's `xttest2` routine (`findit xttest2`).

We have considered one-way fixed effects models, where the effect is attached to the individual. We may also define a two-way fixed effect model, where effects are attached to each unit and time period. 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.

We have considered one-way fixed effects models, where the effect is attached to the individual. We may also define a two-way fixed effect model, where effects are attached to each unit and time period. 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.

The between estimator

Another estimator that may be defined for a panel data set 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, replacing each observation for an individual with their mean behavior. This estimator is not widely used, but has sometimes been applied where the time series data for each individual are thought to be somewhat inaccurate, or when they are assumed to contain random deviations from long-run means. If you assume that the inaccuracy has mean zero over time, a solution to this measurement error problem can be found by averaging the data over time and retaining only one observation per unit.

This could be done explicitly with Stata's `collapse` command. However, you need not form that data set to employ the between estimator, since the command `xtreg` with the `be` (between) option will invoke it. Use of the between estimator requires that $N > K$. Any macro factor that is constant over *individuals* cannot be included in the between estimator, since its average will not differ by individual.

This could be done explicitly with Stata's `collapse` command. However, you need not form that data set to employ the between estimator, since the command `xtreg` with the `be` (between) option will invoke it. Use of the between estimator requires that $N > K$. Any macro factor that is constant over *individuals* cannot be included in the between estimator, since its average will not differ by individual.

We can 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. We might ask, in the context of panel data: 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, whereas the between estimator considers only the latter.

The random effects estimator

As an alternative to considering the individual-specific intercept as a "fixed effect" of that unit, we might consider that the individual effect may be viewed as a random draw from a distribution:

$$y_{it} = X^*_{it}\beta^* + [u_i + \epsilon_{it}] \tag{7}$$

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 or by the between estimator, but that would be inefficient in not taking the nature of the composite disturbance process into account.

A crucial assumption of this model is that u_i is independent of X^*: individual i receives a random draw that gives her a higher wage. That u_i must be independent of individual i's measurable characteristics included among the regressors X^*. If this assumption is not sustained, the random effects estimator will yield inconsistent estimates since the regressors will be correlated with the composite disturbance term.

A crucial assumption of this model is that u_i is independent of X^*: individual i receives a random draw that gives her a higher wage. That u_i must be independent of individual i's measurable characteristics included among the regressors X^*. If this assumption is not sustained, the random effects estimator will yield inconsistent estimates since the regressors will be correlated with the composite disturbance term.

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) as draws from an independent distribution. This greatly reduces the number of parameters to be estimated, and conditional on that independence, allows for inference to be made to the population from which the survey was constructed.

In a large survey, with thousands of individuals, a random effects model will estimate K parameters, whereas a fixed effects model will estimate $(K-1)+N$ parameters, with the sizable loss of $(N-1)$ degrees of freedom. In contrast to fixed effects, the random effects estimator can identify the parameters on time-invariant regressors such as race or gender at the individual level.

In a large survey, with thousands of individuals, a random effects model will estimate K parameters, whereas a fixed effects model will estimate $(K-1)+N$ parameters, with the sizable loss of $(N-1)$ degrees of freedom. In contrast to fixed effects, the random effects estimator can identify the parameters on time-invariant regressors such as race or gender at the individual level.

Therefore, where its use can be warranted, the random effects model is more efficient and allows a broader range of statistical inference. The assumption of the individual effects' independence is testable and should always be tested.

To implement the one-way random effects formulation of Equation (7), we assume that both u and ϵ 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 belonging to the i^{th} unit of the panel, we have the composite error process

$$\eta_{it} = u_i + \epsilon_{it} \tag{8}$$

To implement the one-way random effects formulation of Equation (7), we assume that both u and ϵ 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 belonging to the i^{th} unit of the panel, we have the composite error process

$$\eta_{it} = u_i + \epsilon_{it} \tag{8}$$

This is known as the *error components* model with conditional variance

$$E[\eta_{it}^2 | X^*] = \sigma_u^2 + \sigma_\epsilon^2 \tag{9}$$

and conditional covariance within a unit of

$$E[\eta_{it}\eta_{is} | X^*] = \sigma_u^2, \quad t \neq s. \tag{10}$$

The covariance matrix of these T errors may then be written as

$$\Sigma = \sigma_\epsilon^2 I_T + \sigma_u^2 \iota_T \iota_T'. \tag{11}$$

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

Generalized least squares (GLS) is the estimator for the slope parameters of this model:

$$\begin{aligned} b_{RE} &= (X^{*\prime}\Omega^{-1}X^{*})^{-1}(X^{*\prime}\Omega^{-1}y) \\ &= \left(\sum_i X_i^{*\prime}\Sigma^{-1}X_i^{*}\right)^{-1}\left(\sum_i X_i^{*\prime}\Sigma^{-1}y_i\right) \end{aligned} \tag{12}$$

Generalized least squares (GLS) is the estimator for the slope parameters of this model:

$$\begin{aligned} b_{RE} &= (X^{*\prime}\Omega^{-1}X^{*})^{-1}(X^{*\prime}\Omega^{-1}y) \\ &= \left(\sum_i X_i^{*\prime}\Sigma^{-1}X_i^{*}\right)^{-1}\left(\sum_i X_i^{*\prime}\Sigma^{-1}y_i\right) \end{aligned} \tag{12}$$

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

$$\Sigma^{-1/2} = \sigma_\epsilon^{-1}[I - T^{-1}\theta\iota_T\iota_T'] \tag{13}$$

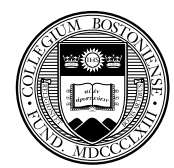
where

$$\theta = 1 - \frac{\sigma_\epsilon}{\sqrt{\sigma_\epsilon^2 + T\sigma_u^2}} \tag{14}$$

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 from each value, we should subtract some fraction of that mean, 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

$$\lambda = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + T\sigma_u^2} = (1 - \theta)^2 \tag{15}$$

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 ϵ_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 is optimal. If $\theta = 1$, $\lambda = 0$ and 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 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 is optimal. If $\theta = 1$, $\lambda = 0$ and 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 GLS random effects estimator applies the optimal λ in the unit interval to the between estimator, whereas the fixed effects estimator arbitrarily imposes $\lambda = 0$. This would only be appropriate if the variation in ϵ was trivial in comparison with the variation in u, since then the indicator variables that identify each unit would, taken together, explain almost all of the variation in the composite error term.

To implement the feasible GLS estimator of the model all we need are consistent estimates of σ_ϵ^2 and σ_u^2. Because the fixed effects model is consistent its residuals can be used to estimate σ_ϵ^2. Likewise, the residuals from the pooled OLS model can be used to 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.

To implement the feasible GLS estimator of the model all we need are consistent estimates of σ_ϵ^2 and σ_u^2. Because the fixed effects model is consistent its residuals can be used to estimate σ_ϵ^2. Likewise, the residuals from the pooled OLS model can be used to 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.

Because the GLS model uses quasi-demeaning, it is capable of including variables that do not vary at the individual level (such as gender or race). Since such variables cannot be included in the LSDV model, an alternative estimator must be defined based on the between estimator's consistent estimate of $(\sigma_u^2 + T^{-1}\sigma_\epsilon^2)$.

The feasible GLS estimator may be executed in Stata using the command `xtreg` with the `re` (random effects) option. The command will display estimates of σ_u^2, σ_ϵ^2 and what Stata calls `rho`: the fraction of variance due to ϵ_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`.

The feasible GLS estimator may be executed in Stata using the command `xtreg` with the `re` (random effects) option. The command will display estimates of σ_u^2, σ_ϵ^2 and what Stata calls `rho`: the fraction of variance due to ϵ_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`.

You can also estimate the parameters of the random effects model with full maximum likelihood. The `mle` option on the `xtreg, re` command requests that estimator. The application of MLE continues to assume that X^* and u are independently distributed, adding the assumption 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.

A Hausman test may be used to test the null hypothesis that the extra orthogonality conditions imposed by the random effects estimator are valid. The fixed effects estimator, which does not impose those conditions, is consistent regardless of the independence of the individual effects. The fixed effects estimates are inefficient if that assumption of independence is warranted. The random effects estimator is efficient under the assumption of independence, but inconsistent otherwise.

A Hausman test may be used to test the null hypothesis that the extra orthogonality conditions imposed by the random effects estimator are valid. The fixed effects estimator, which does not impose those conditions, is consistent regardless of the independence of the individual effects. The fixed effects estimates are inefficient if that assumption of independence is warranted. The random effects estimator is efficient under the assumption of independence, but inconsistent otherwise.

Therefore, we may consider these two alternatives in the Hausman test framework, estimating both models and comparing their common coefficient estimates in a probabilistic sense. If both fixed and random effects models generate consistent point estimates of the slope parameters, they will not differ meaningfully. If the assumption of independence is violated, the inconsistent random effects estimates will differ from their fixed effects counterparts.

To implement the Hausman test, you estimate 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 `fix` 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 independence. 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 fixed effects and random effects vectors of slope coefficients.

The Hausman–Taylor estimator

If the Hausman test indicates that the random effects u_i cannot be considered orthogonal to the individual level error, an instrumental variables estimator may be utilized to generate consistent estimates of the coefficients on the time-invariant variables. The Hausman–Taylor estimator (1981) 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 dividing the regressors into four categories: the interaction of time varying (X) / time invariant (Z) and uncorrelated with u_i (1) / correlated with u_i (2). For example, X_2 are those time-varying regressors that are thought to be correlated with u_i. Identification of the parameters requires that K_1 (the number of X_1 variables) be at least as large as L_2 (the number of Z_2 variables).

Their approach is based on dividing the regressors into four categories: the interaction of time varying (X) / time invariant (Z) and uncorrelated with u_i (1) / correlated with u_i (2). For example, X_2 are those time-varying regressors that are thought to be correlated with u_i. Identification of the parameters requires that K_1 (the number of X_1 variables) be at least as large as L_2 (the number of Z_2 variables).

The application of the Hausman–Taylor 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.

The IV estimator for panel data

Stata also provides an instrumental variables estimator for the fixed effects and random effects models in which some of the X variables are correlated with the idiosyncratic error ϵ. These are quite different assumptions about the nature of any suspected correlation between regressor and the composite error term from those underlying the Hausman–Taylor estimator. The `xtivreg` command also supports fixed effects, between effects, and first-differenced estimators in an instrumental variables context.

Considering our discussion of instrumental variables estimation via `ivreg2`, the features of `ivreg2` are also available for panel data in `xtivreg2`, which is a "wrapper" for `ivreg2`. This routine of Mark Schaffer's extends Stata's `xtivreg`'s support for the fixed effect (`fe`) and first difference (`fd`) estimators. The `xtivreg2` routine is available from `ssc`.

Considering our discussion of instrumental variables estimation via `ivreg2`, the features of `ivreg2` are also available for panel data in `xtivreg2`, which is a "wrapper" for `ivreg2`. This routine of Mark Schaffer's extends Stata's `xtivreg`'s support for the fixed effect (`fe`) and first difference (`fd`) estimators. The `xtivreg2` routine is available from `ssc`.

Just as `ivreg2` may be used to conduct a Hausman test of IV vs. OLS, Schaffer and Stillman's `xtoverid` routine may be used to conduct a Hausman test of random effects vs. fixed effects after `xtreg, re` and `xtivreg, re`. This routine can also calculate tests of overidentifying restrictions after those two commands as well as `xthtaylor`. The `xtoverid` routine is also available from `ssc`.

The first difference estimator

The within transformation used by fixed effects models removes unobserved heterogeneity at the unit level. The same can be achieved by first differencing the original equation (which removes the constant term). In fact, if $T = 2$, the fixed effects and first difference estimates are identical. For $T > 2$, the effects will not be identical, but they are both consistent estimators of the original model. Stata's `xtreg` does not provide the first difference estimator, but `xtivreg2` provides this option as the `fd` model.

The first difference estimator

The within transformation used by fixed effects models removes unobserved heterogeneity at the unit level. The same can be achieved by first differencing the original equation (which removes the constant term). In fact, if $T = 2$, the fixed effects and first difference estimates are identical. For $T > 2$, the effects will not be identical, but they are both consistent estimators of the original model. Stata's `xtreg` does not provide the first difference estimator, but `xtivreg2` provides this option as the `fd` model.

The ability of first differencing to remove unobserved heterogeneity also underlies the family of estimators that have been developed for dynamic panel data (DPD) models. These models contain one or more lagged dependent variables, allowing for the modeling of a partial adjustment mechanism.

A serious difficulty arises with the one-way fixed effects model in the context of a *dynamic panel data* (DPD) model 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.

A serious difficulty arises with the one-way fixed effects model in the context of a *dynamic panel data* (DPD) model 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.

The demeaning operation creates a regressor which *cannot* be distributed independently of the error term. Nickell demonstrates that the inconsistency of $\hat{\rho}$ as $N \rightarrow \infty$ 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 $(\hat{\rho} - \rho)$ as $N \rightarrow \infty$ 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.

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:

$$y_{it} = \beta_1 + \rho y_{i,t-1} + X_{it}\beta_2 + \ u_i + \epsilon_{it} \tag{16}$$

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

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta X_{it}\beta_2 + \Delta \epsilon_{it} \tag{17}$$

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.

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).

Dynamic panel data estimators

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 Generalized Method of Moments (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

$$\begin{aligned} y_{it} &= X_{it}\beta_1 + W_{it}\beta_2 + v_{it} \\ v_{it} &= u_i + \epsilon_{it} \end{aligned} \tag{18}$$

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 more general version, allowing for autocorrelated errors, is available as `xtdpd`.

A potential weakness in the Arellano–Bond *DPD* estimator was revealed in later work by Arellano and Bover (1995) and Blundell and Bond (1998). 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.

A potential weakness in the Arellano–Bond *DPD* estimator was revealed in later work by Arellano and Bover (1995) and Blundell and Bond (1998). 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. This estimator is available in Stata as `xtdpdsys`.

An excellent alternative to Stata's built-in commands is David Roodman's `xtabond2`, available from SSC (`findit xtabond2`). It is very well documented in his paper, referenced above. The `xtabond2` routine handles both the difference and system GMM estimators and provides several additional features—such as the orthogonal deviations transformation—not available in official Stata's commands.

An excellent alternative to Stata's built-in commands is David Roodman's `xtabond2`, available from SSC (`findit xtabond2`). It is very well documented in his paper, referenced above. The `xtabond2` routine handles both the difference and system GMM estimators and provides several additional features—such as the orthogonal deviations transformation—not available in official Stata's commands.

As any of the DPD estimators are instrumental variables methods, it is particularly important to evaluate the Sargan–Hansen test results when they are applied. Roodman's `xtabond2` provides *C* tests (as discussed in re `ivreg2`) for groups of instruments. In his routine, instruments can be either "GMM-style" or "IV-style". The former are constructed per the Arellano–Bond logic, making use of multiple lags; the latter are included as is in the instrument matrix. For the system GMM estimator (the default in `xtabond2` instruments may be specified as applying to the differenced equations, the level equations or both.

Another important diagnostic in DPD estimation is the *AR* test for autocorrelation of the residuals. By construction, the residuals of the differenced equation should possess serial correlation, but if the assumption of serial independence in the original errors is warranted, the differenced residuals should not exhibit significant $AR(2)$ behavior. These statistics are produced in the `xtabond` and `xtabond2` output. If a significant $AR(2)$ statistic is encountered, the second lags of endogenous variables will not be appropriate instruments for their current values.

Another important diagnostic in DPD estimation is the *AR* test for autocorrelation of the residuals. By construction, the residuals of the differenced equation should possess serial correlation, but if the assumption of serial independence in the original errors is warranted, the differenced residuals should not exhibit significant $AR(2)$ behavior. These statistics are produced in the `xtabond` and `xtabond2` output. If a significant $AR(2)$ statistic is encountered, the second lags of endogenous variables will not be appropriate instruments for their current values.

A useful feature of `xtabond2` is the ability to specify, for GMM-style instruments, the limits on how many lags are to be included. If T is fairly large (more than 7–8) an unrestricted set of lags will introduce a huge number of instruments, with a possible loss of efficiency. By using the lag limits options, you may specify, for instance, that only lags 2–5 are to be used in constructing the GMM instruments.

Although the DPD estimators are linear estimators, they are highly sensitive to the particular specification of the model and its instruments. There is no substitute for experimentation with the various parameters of the specification to ensure that your results are reasonably robust to variations in the instrument set and lags used. If you are going to work with DPD models, you should study Roodman's "How to do `xtabond2`" paper so that you fully understand the nuances of this estimation strategy.