Using Arellano – Bond Dynamic Panel GMM Estimators in Stata

Tutorial with Examples using Stata 9.0
(xtabond and xtabond2)

Elitza Mileva,
Economics Department
Fordham University

July 9, 2007

1. The model

The following model examines the impact of capital flows on investment in a panel dataset of 22 countries for 10 years (1995 – 2004):

$$I_{it} = \beta_1 I_{i,t-1} + \beta_2 K_{it} + \beta_3 X_{it} + u_{it}. \qquad (1)$$

In equation (1) above I_{it} is gross fixed capital formation as a percentage of GDP and I_{it-1} is its lagged value. K_{it} is a matrix of the components of foreign resource flows – FDI, loans and portfolio (equity and bonds) – as percentage shares of GDP. X_{it} is a matrix of the following control variables: lagged real GDP growth to account for the accelerator effect; the absolute value of one step ahead growth forecast errors as a measure of uncertainty; the change in the log terms of trade to gauge the price of imported capital goods; and, finally, the deviation of M2 from its three-year trend as a proxy for the liquidity available to finance investment.

2. Why the Arellano – Bond GMM estimator?

Several econometric problems may arise from estimating equation (1):

1. The capital flows variables in K_{it} are assumed to be endogenous. Because causality may run in both directions – from capital inflows to investment and vice versa – these regressors may be correlated with the error term.

2. Time-invariant country characteristics (fixed effects), such as geography and demographics, may be correlated with the explanatory variables. The fixed effects are contained in the error term in equation (1), which consists of the unobserved country-specific effects, v_i, and the observation-specific errors, e_{it}:

$$u_{it} = v_i + e_{it} \qquad (2).$$

3. The presence of the lagged dependent variable I_{it-1} gives rise to autocorrelation.

4. The panel dataset has a short time dimension ($T = 10$) and a larger country dimension ($N = 22$).

To solve **problem 1** (and problem 2) one would usually use fixed-effects instrumental variables estimation (two-stage least squares or 2SLS), which is what I tried first. The exogenous instruments I used were the following: the aggregate long-term capital inflows to the countries in our sample as a group as a percentage of the sum of their cumulative GDP (I labelled these 'regional flows'), an index of financial openness and the EBRD transition index. However, the first-stage statistics of the 2SLS regressions showed that my instruments were weak. With weak instruments the fixed-effects IV estimators are likely to be biased in the way of the OLS estimators. Therefore, I decided to use the Arellano – Bond (1991) difference GMM estimator first proposed by Holtz-Eakin, Newey and Rosen (1988). Instead of using only the exogenous instruments listed above lagged levels of the endogenous regressors in K_{it} (FDI, loans and portfolio) are also added. This makes the endogenous variables pre-determined and, therefore, not correlated with the error term in equation (1).

To cope with **problem 2** (fixed effects) the difference GMM uses first-differences to transform equation (1) into

$$\Delta I_{it} = \beta_1 \Delta I_{i,t-1} + \beta_2 \Delta K_{it} + \beta_3 \Delta X_{it} + \Delta u_{it} \qquad (3).$$

(In general form the transformation is given by: $\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta x'_{it}\beta + \Delta u_{it}$.)

By transforming the regressors by first differencing the fixed country-specific effect is removed, because it does not vary with time. From equation (2) we get

$$\Delta u_{it} = \Delta v_i + \Delta e_{it}$$

or

$$u_{it} - u_{i,t-1} = (v_i - v_i) + (e_{it} - e_{i,t-1}) = e_{it} - e_{i,t-1}.$$

The first-differenced lagged dependent variable (**problem 3**) is also instrumented with its past levels.

Finally, the Arellano – Bond estimator was designed for small-T large-N panels (**problem 4**). In large-T panels a shock to the country's fixed effect, which shows in the error term, will decline with time. Similarly, the correlation of the lagged dependent variable with the error term will be insignificant (see Roodman, 2006). In these cases, one does not necessarily have to use the Arellano – Bond estimator.

3. Using the Arellano – Bond difference GMM estimator in Stata

3.1 Import data into Stata

The easiest way to get panel data into Stata is to organize your Excel spreadsheet in the following way:

ctry	ctry_dum	year	inv	growth	uncert	tot	dev_m2	fin_integr	trans_index	fdi	loans	portfolio	flows_eeca
ALB	1	1995	18.000	8.900	8.444	0.215	.	3.000	2.333	0.861	-0.005	0.000	1.121
ALB	1	1996	21.044	9.100	6.614	-0.112	.	3.000	2.519	0.994	0.050	0.000	1.198
ALB	1	1997	16.829	-10.200	12.247	0.057	9.447	3.000	2.519	0.580	-0.013	0.000	1.783
ALB	1	1998	16.296	12.700	16.874	0.019	3.281	3.024	2.519	0.480	-0.019	0.000	2.365
ALB	1	1999	20.005	10.100	6.783	0.071	1.444	3.024	2.557	0.389	-0.035	0.000	1.826
ALB	1	2000	24.736	7.300	3.750	-0.006	2.572	3.024	2.778	1.245	-0.009	0.000	1.488
ALB	1	2001	29.215	7.200	4.023	-0.018	2.654	3.024	2.814	1.648	-0.031	0.000	1.263
ALB	1	2002	26.156	3.400	0.045	-0.010	2.937	3.024	2.814	1.002	0.005	0.000	1.718
ALB	1	2003	25.013	6.002	3.716	0.011	-0.455	3.024	2.814	1.225	-0.019	0.000	1.894
ALB	1	2004	23.686	5.900	2.543	0.040	-1.298	3.024	2.889	2.701	0.188	0.000	3.288
ARM	2	1995	16.154	6.900	17.601	-0.103	-10.808	2.000	2.112	0.394	0.000	0.000	1.121
ARM	2	1996	17.885	5.865	7.872	0.302	0.261	3.000	2.444	0.309	0.000	0.033	1.198

…

Note that all observations (i.e. country 1 period 1; country 1 period 2; etc.) are stacked vertically and the variable are listed horizontally.

Save the Excel worksheet as a text file (.txt, .csv, etc.). Open Stata and import the data by choosing File, Import, ASCII data created by spreadsheet, and click on the Browse button. Alternatively, you can type the following command in the command window, if your text file is located on the C drive:

```
insheet using "C:\ABExampleData.txt"
(14 vars, 220 obs)
```

(Note that from now on text in blue will show Stata commands or their components.)

3.2 Set the dataset as a panel

Next, save your dataset as a panel by selecting Statistics, Longitudinal / Panel data, Setup & Utilities, Declare dataset to be cross-sectional time series. Choose a variable that identifies the time dimension (year, in this example) and a variable that identifies the panel ID (ctry_dum, in this example). Stata needs a numerical variable for the panel ID so the variable ctry, which is a string variable, won't work. Alternatively, you can type the following command:

```
tsset ctry_dum year
       panel variable:  ctry_dum (strongly balanced)
        time variable:  year, 1995 to 2004
```

3.3 Stata command: xtabond

Two Arellano–Bond estimators are available for Stata 9.0 – one incorporated into Stata 9 (called xtabond) and one proprietor program written by Roodman (2006) (called xtabond2). First is discussed the former (Stata 10.0 will have two AB estimators built in, including it version of the system estimator).

Click on Statistics, Longitudinal / Panel data, Dynamic panel data, Arellano – Bond regression (RE). Stata displays a window, in which you can easily select the dependent variable, the endogenous and exogenous independent variables as well as the lags of the instruments.

3.4 Stata command: xtabond2

Although the above-mentioned Stata menu option is easier to use, I have found Roodman's proprietary program (xtabond2) better – it is more flexible and has a better help file and "how to do xtabond2" paper (see in the references). xtabond2 can do everything that xtabond does and has many additional features. See the Stata help file or the paper for a description of the improvements offered by Roodman's program. The disadvantage of xtabond2 is that you actually have to type the program code – there is no menu for it.

Since xtabond2 is not an official command of Stata 9, it has to be downloaded from the Internet http://ideas.repec.org/c/boc/bocode/s435901.html or by typing the following command:

```
ssc install xtabond2
```

If you have to download all xtabond2-related files from the *repec* website, make sure you save each file in the appropriate ado folder in your Stata folder, that is in the folder of the first letter of the file name as it is listed on the website.

(xtabond2 may be directly available with Stata 10, or it may include a different system routine)

The following command shows you the help file:

```
help xtabond2
```

Below is the command I used to estimate equation (1) followed by the Stata output:

```
xtabond2 inv l.inv fdi loans portfolio l.growth uncert tot dev_m2, gmm (inv fdi
loans portfolio, lag (2 2)) iv(fin_integr trans_index flows_eeca l.growth
uncert tot dev_m2) nolevel small
Favoring space over speed. To switch, type or click on mata: mata set matafavor
speed, perm.
Warning: Number of instruments may be large relative to number of observations.
Suggested rule of thumb: keep number of instruments <= number of groups.

Arellano-Bond dynamic panel-data estimation, one-step difference GMM results
------------------------------------------------------------------------------
Group variable: ctry_dum                        Number of obs      =       165
Time variable : year                            Number of groups   =        22
Number of instruments = 39                      Obs per group: min =         3
F(8, 157)     =      6.88                                      avg =      7.50
Prob > F      =     0.000                                      max =         8
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inv |
         L1. |   .2922856    .111738     2.62   0.010     .0715819    .5129893
         fdi |   .5202847   .2094545     2.48   0.014     .1065725     .933997
       loans |   .2789421   .1638248     1.70   0.091     -.044643    .6025271
   portfolio |  -.0086876   .3376843    -0.03   0.980    -.6756779    .6583028
      growth |
         L1. |   .1167961   .0555715     2.10   0.037     .0070319    .2265604
      uncert |   .0397982   .0673439     0.59   0.555    -.0932187     .172815
         tot |   .9193659   1.916147     0.48   0.632    -2.865388    4.704119
      dev_m2 |   .0443079   .0760188     0.58   0.561    -.1058435    .1944594
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(31) =  36.42    Prob > chi2 =  0.231

Arellano-Bond test for AR(1) in first differences: z =  -0.01  Pr > z =  0.992
Arellano-Bond test for AR(2) in first differences: z =  -0.48  Pr > z =  0.628
```

As you can see, the command xtabond2 is followed by the dependent variable (inv) and the list of all right-hand-side variables:

```
xtabond2 inv l.inv fdi loans portfolio l.growth uncert tot dev_m2
```

The lag operator is given by l. as in l.inv or l2.inv for 2 lags of inv.

After the comma are given two lists of variables. gmm( ) (or gmmstyle( )) lists the endogenous variables, which are instrumented with GMM-style instruments, i.e. lagged values of the variables in levels:

```
gmm (inv fdi loans portfolio, lag (2 2))
```

With lag (2 2) I have instructed Stata to use only the second lag of the endogenous variables as instruments. Due to the small number of countries in my sample a large number of instruments causes the Sargan test (explained below) to be weak. The rule of thumb is to keep the number of instruments less than or equal to the number of groups. Stata warns you about that at the top of the output table. The second lag is required, because it is not correlated with the current error term, while the first lag is. Generally, one can experiment with a second or deeper lags to find a good instrument, but using deeper lags reduces sample size. If the number of countries is large enough, one may use all available lags (second and deeper lags) as instruments.

The second list of explanatory variables, iv ( ) (or ivstyle ( )), lists all strictly exogenous variables (l.growth, uncert, tot, dev_m2) as well as the additional instrumental variables (fin_integr, trans_index, flows_eeca), which are not part of equation (1) and, therefore, are not listed before the comma in the Stata command. What this option essentially does for the included exogenous variables is tell Stata to use the variables themselves as their own instruments.

```
iv(fin_integr trans_index flows_eeca l.growth uncert tot dev_m2)
```

Growth is lagged in this case due to economic theory and not because it is required by the regression. Another advantage of xtabond2 is that it actually allows you to use lag operators in the instruments matrix, while xtabond does not.

nolevel (or noleveleq) tells Stata to apply the difference GMM estimator. By default xtabond2 will apply the system GMM, if you don’t specify nolevel. (System GMM is discussed next.)

small tells Stata to use the small-sample adjustment and report t- instead of z-statistics and the Wald chi-squared test instead of the F test.

Stata offers additional options not shown in the example above:

twostep specifies that the two-step estimator is calculated instead of the default one-step. In two-step estimation, the standard covariance matrix is robust to panel-specific autocorrelation and heteroskedasticity, but the standard errors are downward biased. Use twostep robust to get the finite-sample corrected two-step covariance matrix.

robust specifies that the resulting standard errors are consistent with panel-specific autocorrelation and heteroskedasticity in one-step estimation.

By default Stata also reports three additional tests: Sargan test, AR(1) and AR(2) tests.

The Sargan test has a null hypothesis of "the instruments as a group are exogenous". Therefore, the higher the p-value of the Sargan statistic the better. In robust estimation Stata reports the Hansen J statistic instead of the Sargan with the same null hypothesis.

The Arellano – Bond test for autocorrelation has a null hypothesis of no autocorrelation and is applied to the differenced residuals. The test for AR (1) process in first differences usually rejects the null hypothesis (though not in my example), but this is expected since

$\Delta e_{it} = e_{it} - e_{i,t-1}$ and $\Delta e_{i,t-1} = e_{i,t-1} - e_{i,t-2}$ both have $e_{i,t-1}$.

The test for AR (2) in first differences is more important, because it will detect autocorrelation in levels.

Before closing Stata you can save the data file in .dta format, which is the Stata data format. Choose File, Save As or type:

```
save "C:\ABExample.dta"
```

When you open that file next time, all settings, such as the panel-data setting, or any new variables you have created will be saved.

4. Using the Arellano – Bond system GMM estimator in Stata

Sometimes the lagged levels of the regressors are poor instruments for the first-differenced regressors. In this case, one should use the augmented version – "system GMM". The system GMM estimator uses the levels equation (e.g. equation (1) in this example) to obtain a system of two equations: one differenced and one in levels. By adding the second equation additional instruments can be obtained. Thus the variables in levels in the second equation are instrumented with their own first differences. This usually increases efficiency.

Below is the command and Stata output for Arellano – Bond System GMM estimator. Note that nolevel no longer is included after the comma in the command and Stata defaults to the system GMM. Including the equation in levels does not difference out the constant, therefore, if the model does not call for a constant, type noconst after the comma in the command.

```
xtabond2 inv l.inv fdi loans portfolio l.growth uncert tot dev_m2, gmm (inv fdi
loans portfolio, lag (3 3)) iv(fin_integr trans_index flows_eeca l.growth
uncert tot dev_m2) small noconst
```

```
Favoring space over speed. To switch, type or click on mata: mata set matafavor
speed, perm.
Warning: Number of instruments may be large relative to number of observations.
Suggested rule of thumb: keep number of instruments <= number of groups.

Arellano-Bond dynamic panel-data estimation, one-step system GMM results
------------------------------------------------------------------------------
Group variable: ctry_dum                        Number of obs      =       187
Time variable : year                            Number of groups   =        22
Number of instruments = 63                      Obs per group: min =         4
F(8, 179)     =   1700.28                                      avg =      8.50
Prob > F      =     0.000                                      max =         9
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inv |
         L1. |    .898973    .025188    35.69   0.000     .8492693    .9486767
         fdi |   .9096135   .2207568     4.12   0.000     .4739929    1.345234
       loans |   .1813443   .1994594     0.91   0.364    -.2122499    .5749386
   portfolio |   -.697416   .4072526    -1.71   0.089     -1.50105    .1062178
      growth |
         L1. |   .1028205   .0514219     2.00   0.047     .0013493    .2042917
      uncert |   .1431728   .0564829     2.53   0.012     .0317148    .2546307
         tot |  -2.131275   2.446291    -0.87   0.385    -6.958554    2.696004
      dev_m2 |   .0076649    .081528     0.09   0.925    -.1532147    .1685446
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(55) =  43.77    Prob > chi2 =  0.862

Arellano-Bond test for AR(1) in first differences: z =  -1.88  Pr > z =  0.061
Arellano-Bond test for AR(2) in first differences: z =  -0.86  Pr > z =  0.391
```

As the output table above shows, using system GMM increased efficiency. There are, however, two important points to be made about using system GMM. First, because system GMM uses more instruments than the difference GMM it may not be appropriate to use system GMM with a dataset with a small number of countries. Recall that when the number of instruments is greater than the number of countries the Sargan test may be weak.

Second, in a panel with fixed effects including the equation in levels requires a new assumption – the first-differenced instruments used for the variables in levels should not be correlated with the unobserved country effects. Roodman (2006) discusses how this assumption depends on assumptions about the initial conditions. Some authors prefer to include in the levels equation only those variables, which are uncorrelated with the fixed effects. xtabond2 offers the equation () sub-option, which specifies which equation should use the instruments: first-difference only (equation (diff)) or levels only (equation (level)). The default is both equations.

In our example (and only for the sake of understanding the Stata commands) we may decide that only inv, fdi, loans and portfolio are correlated with the unobserved country effects while all other instruments are not. In this case the Stata command and output are the following:

```
xtabond2 inv l.inv fdi loans portfolio l.growth uncert tot dev_m2, gmm (inv fdi
loans portfolio, eq(diff) lag (3 3)) iv(fin_integr trans_index flows_eeca
l.growth uncert tot dev_m2) small noconst
Favoring space over speed. To switch, type or click on mata: mata set matafavor
speed, perm.
Warning: Number of instruments may be large relative to number of observations.
Suggested rule of thumb: keep number of instruments <= number of groups.

Arellano-Bond dynamic panel-data estimation, one-step system GMM results
------------------------------------------------------------------------------
Group variable: ctry_dum                        Number of obs      =       187
Time variable : year                            Number of groups   =        22
Number of instruments = 35                      Obs per group: min =         4
F(8, 179)     =   1560.72                                      avg =      8.50
Prob > F      =     0.000                                      max =         9
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inv |
         L1. |   .8880598    .033887    26.21   0.000     .8211904    .9549292
         fdi |   .9775032   .3773267     2.59   0.010     .2329224    1.722084
       loans |   .5800662   .3150042     1.84   0.067    -.0415332    1.201666
   portfolio |  -.8037535   .5324702    -1.51   0.133     -1.85448    .2469728
      growth |
         L1. |   .0911793   .0573559     1.59   0.114    -.0220013    .2043599
      uncert |   .1231768   .0639041     1.93   0.055    -.0029254    .2492791
         tot |  -.9632624   2.619933    -0.37   0.714     -6.13319    4.206666
      dev_m2 |  -.0082295   .0876145    -0.09   0.925    -.1811197    .1646608
------------------------------------------------------------------------------
Sargan test of overid. restrictions: chi2(27) =  18.91    Prob > chi2 =  0.873

Arellano-Bond test for AR(1) in first differences: z =  -1.80  Pr > z =  0.073
Arellano-Bond test for AR(2) in first differences: z =  -0.87  Pr > z =  0.383
------------------------------------------------------------------------------
```

5. References

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

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

Roodman, D. (December 2006). How to do xtabond2: an introduction to “Difference” and “System” GMM in Stata. *Center for Global Development Working Paper Number 103*.