

# Econometric analysis of dynamic panel-data models using Stata

David M. Drukker

StataCorp

Summer North American Stata Users Group meeting  
July 24-25, 2008

- 1 Dynamic panel-data models
- 2 The Arellano-Bond estimator
- 3 The Arellano-Bover/Blundell-Bond estimator

# Introduction

- We are interested in estimating the parameters of models of the form

$$y_{it} = y_{it-1}\gamma + \mathbf{x}_{it}\beta + u_i + \epsilon_{it}$$

for  $i = \{1, \dots, N\}$  and  $t = \{1, \dots, T\}$  using datasets with large  $N$  and fixed  $T$

- By construction,  $y_{it-1}$  is correlated with the unobserved individual-level effect  $u_i$
- Removing  $u_i$  by the within transform (removing the panel-level means) produces an inconsistent estimator with  $T$  fixed
- First difference both sides and look for instrumental-variables (IV) and generalized method-of-moments (GMM) estimators

# The Arellano-Bond estimator I

- First differencing the model equation yields

$$\Delta y_{it} = \Delta y_{it-1}\gamma + \Delta \mathbf{x}_{it}\beta + \Delta \epsilon_{it}$$

- The  $u_i$  are gone, but the  $y_{it-1}$  in  $\Delta y_{it-1}$  is a function of the  $\epsilon_{it-1}$  which is also in  $\Delta \epsilon_{it}$   
So  $\Delta y_{it-1}$  is correlated with  $\Delta \epsilon_{it}$  by construction
- [Anderson and Hsiao(1981)] suggested a 2SLS estimator based on further lags of  $\Delta y_{it}$  as instruments for  $\Delta y_{it-1}$ 
  - For instance, if  $\epsilon_{it}$  is IID over  $i$  and  $t$ ,  $\Delta y_{it-2}$  would be a valid instrument for  $\Delta y_{it-1}$
- [Anderson and Hsiao(1981)] also suggested a 2SLS estimator based on lagged levels of  $\Delta y_{it}$  as instruments for  $\Delta y_{it-1}$ 
  - For instance, if  $\epsilon_{it}$  is IID over  $i$  and  $t$ ,  $y_{it-2}$  would be a valid instrument for  $\Delta y_{it-1}$

# The Arellano-Bond estimator II

- [Holtz-Eakin et al.(1988)Holtz-Eakin, Newey, and Rosen] and [Arellano and Bond(1991)] showed how to construct estimators based on moment equations constructed from further lagged levels of  $y_{it}$  and the first-differenced errors
  - We are creating moment conditions using lagged levels of the dependent variable with first differences of the errors  $\epsilon_{it}$
  - First-differences of strictly exogenous covariates are also used to create moment conditions
- Assume that  $\epsilon_{it}$  are IID over  $i$  and  $t$ , i.e. no serial correlation in the errors
  - We will drop this assumption later
- We have more instruments than parameters, so use GMM framework

# Strict Exogeneity

- If the regressors are strictly exogenous,  $\epsilon_{it}$  cannot affect  $\mathbf{x}_{is}$  for any  $s$  or  $t$
- If the regressors are predetermined,  $\epsilon_{it}$  may affect  $\mathbf{x}_{is}$  for  $s > t$
- Dynamic panel-data estimators allow for predetermined regressors

# How strict is strict exogeneity?

- Strict exogeneity rules out any feedback from the idiosyncratic shock at time  $t$  to a regressor at time  $s > t$
- Consider the following model

$$\ln(\text{income}_{it}) = \alpha + \text{education}_{it}\beta_1 + \text{married}_{it}\beta_2 + \nu_i + \epsilon_{it}$$

- where we are modeling the log of income as a function of years of education and an indicator for whether or not person  $i$  is married at time  $t$
- Strict exogeneity requires that  $\epsilon_{it}$  be unrelated to  $\text{married}_{is}$  for  $s > t$ , which rules out negative-economic shocks from causing divorces in the future

# The Arellano-Bond estimator III

- The moment conditions formed by assuming that particular lagged levels of the dependent variable are orthogonal to the differenced disturbances are known as GMM-type moment conditions
  - Sometimes they are called sequential moment conditions
- The moment conditions formed using the strictly exogenous covariates are just standard IV moment conditions, so they are called standard moment conditions
- The dynamic panel-data estimators in Stata report which transforms of which variables were used as instruments

# GMM

- In GMM estimators, we weight the vector of sample-average moment conditions by the inverse of a positive definite matrix
- When that matrix is the covariance matrix of the moment conditions, we have an efficient GMM estimator
- In the case of nonidentically distributed disturbances, we can use a two-step GMM estimator that estimates the covariance matrix of the moment conditions using the first-step residuals
- Although the large-sample robust variance-covariance matrix of the two-step estimator does not depend on the fact that estimated residuals were used, simulation studies have found that that Windmeijer's bias-corrected estimator performs much better

## xtabond

- `xtabond depvar indepvars [if] [in], [, options]`

```
. use dpdcrime
. describe
Contains data from dpdcrime.dta
  obs:          8,000
  vars:          7                24 May 2008 17:44
  size:         416,000 (99.2% of memory free)  (_dta has notes)
```

variable name	storage type	display format	value label	variable label
id	float	%9.0g		
t	float	%9.0g		
policepc	double	%10.0g		police officers per thousand
arrestp	double	%10.0g		arrests/crimes
convictp	double	%10.0g		convictions/arrests
legalwage	double	%10.0g		legal wage index 0-1 scale
crime	double	%10.0g		property-crime index 0-50 scale

```
Sorted by:  id  t
```

## xtabond II

```
. xtabond crime legalwage policepc, nocons
Arellano-Bond dynamic panel-data estimation Number of obs      =      6000
Group variable: id                          Number of groups     =      1000
Time variable: t

Obs per group:   min =      6
                  avg =      6
                  max =      6

Number of instruments =      23          Wald chi2(3)         =    15463.20
                                          Prob > chi2          =      0.0000
```

## One-step results

crime	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
crime						
L1.	.6517166	.011831	55.09	0.000	.6285283	.674905
legalwage	-.7046974	.0272204	-25.89	0.000	-.7580483	-.6513464
policepc	-1.657297	.0178262	-92.97	0.000	-1.692236	-1.622358

```
Instruments for differenced equation
GMM-type: L(2/.)crime
Standard: D.legalwage D.policepc
. estimates store ab1
```

# Where did all the instruments come from?

- legalwage policepc are modeled as strictly exogenous, and each contribute 1 instrument
- The remaining 21 instruments come from the  $p - 2$  instruments available in periods  $p = 3, 4, 5, 6, 7, 8$ 
  - In period 3,  $y_{i1}$  is a valid instrument for  $\Delta y_{i3}$
  - In period 4,  $y_{i1}$  and  $y_{i2}$  are valid instruments for  $\Delta y_{i4}$
  - In period 5,  $y_{i1}$ ,  $y_{i2}$  and  $y_{i3}$  are valid instruments for  $\Delta y_{i5}$
  - In period 6,  $y_{i1}$ ,  $y_{i2}$ ,  $y_{i3}$  and  $y_{i4}$  are valid instruments for  $\Delta y_{i6}$
  - and so on
- So in a model with one lag of the dependent variable,  $k$  strictly exogenous variables and  $p = T - 2$  periods from which to form moment equations, there are  $k + p * (p + 1)/2$  moment conditions.
  - $2 + 6 * 7/2 = 23$  for the example above

# Postestimation specification tests

- Use `estat sargan` to get the Sargan test of the null hypothesis that model and overidentifying conditions are correct specified

```
. quietly xtabond crime legalwage policepc, nocons
. estimates store ab1
. estat sargan
Sargan test of overidentifying restrictions
H0: overidentifying restrictions are valid
chi2(20) = 46.05784
Prob > chi2 = 0.0008
```

- Use `estat abond` to get the Arellano-Bond test that there is no serial correlation in the first-differenced disturbances

```
. estat abond
Arellano-Bond test for zero autocorrelation in first-differenced errors
```

Order	z	Prob > z
1	-15.906	0.0000
2	-3.4158	0.0006

H0: no autocorrelation

# The two-step estimator

```
. xtabond crime legalwage policepc, nocons twostep
Arellano-Bond dynamic panel-data estimation Number of obs      =      6000
Group variable: id                Number of groups       =      1000
Time variable: t

                                Obs per group:   min =         6
                                                avg =         6
                                                max =         6

Number of instruments =      23                Wald chi2(3)          =      8739.60
                                                Prob > chi2           =      0.0000
```

## Two-step results

crime	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
crime						
L1.	.6509223	.0220091	29.58	0.000	.6077853	.6940594
legalwage	-.7079127	.0269015	-26.32	0.000	-.7606386	-.6551868
policepc	-1.66315	.0273474	-60.82	0.000	-1.71675	-1.60955

Warning: gmm two-step standard errors are biased; robust standard errors are recommended.

Instruments for differenced equation

GMM-type: L(2/.)crime

Standard: D.legalwage D.policepc

# The two-step estimator with robust standard-errors

- Specifying `vce(robust)` produces an estimated VCE that is robust to heteroskedasticity
- There is a result in the large-sample theory for GMM which states that the VCE of the two-step estimator does not depend on the fact that it uses the residuals from the first step
- For some problems, simulation studies have found that the IID and robust large-sample estimators of the VCE of the two-step GMM estimator have large finite-sample biases

These finite-sample bias cause large differences between the size and rejection rates of Wald tests

- [Windmeijer(2005)] derives an estimate of this finite-sample bias and uses it to bias correct the robust estimator of the VCE of the two-step GMM estimator

## The two-step estimator with robust standard-errors II

```
. xtabond crime legalwage policepc, nocons twostep vce(robust)
Arellano-Bond dynamic panel-data estimation   Number of obs   =   6000
Group variable: id                           Number of groups =   1000
Time variable: t

Obs per group:   min =   6
                  avg =   6
                  max =   6

Number of instruments =   23                Wald chi2(3)     =   7958.93
                                                Prob > chi2      =   0.0000
```

## Two-step results

crime	Coef.	WC-Robust Std. Err.	z	P> z	[95% Conf. Interval]	
crime						
L1.	.6509223	.0235159	27.68	0.000	.6048321	.6970126
legalwage	-.7079127	.0277065	-25.55	0.000	-.7622164	-.653609
policepc	-1.66315	.0286174	-58.12	0.000	-1.719239	-1.607061

```
Instruments for differenced equation
GMM-type: L(2/.)crime
Standard: D.legalwage D.policepc
```

# The two-step estimator with robust standard-errors II

- The distribution of the Sargan test is not known when the disturbances are heteroskedastic, so `estat sargan` is not available after specifying `vce(robust)`
- A robust version of the Arellano-Bond test for serial correlation is produced after specifying `vce(robust)`

```
. estat abond
```

```
Arellano-Bond test for zero autocorrelation in first-differenced errors
```

Order	z	Prob > z
1	-12.269	0.0000
2	-2.9051	0.0037

```
H0: no autocorrelation
```

# Predetermined variables

- Thus far we assumed that the variables in  $\mathbf{x}_{it}$  are strictly exogenous, i.e.  $E[\mathbf{x}_{is}\epsilon_{it}] = \mathbf{0}$  for all  $s$  and  $t$
- If instead, if we have  $E[\mathbf{x}_{is}\epsilon_{it}] = \mathbf{0}$  for  $s \leq t$  but allow  $E[\mathbf{x}_{is}\epsilon_{it}] \neq \mathbf{0}$  for  $s > t$ , the variables are said to be predetermined
  - Suppose that a positive shock to the crime rate today caused an increase in the police per capita tomorrow
- When the variables are predetermined, it means that we cannot include the whole vector of differences of observed  $\mathbf{x}_{it}$  into the instrument matrix
- We just include the levels of  $\mathbf{x}_{it}$  for those time periods that are assumed to be unrelated to  $\Delta\epsilon_{it}$

# Predetermined variables II

```
. xtabond crime legalwage , twostep vce(robust) pre(policepc)
Arellano-Bond dynamic panel-data estimation Number of obs      =      6000
Group variable: id                               Number of groups    =      1000
Time variable: t
Obs per group:  min =          6
                  avg =          6
                  max =          6
Number of instruments =      50                Wald chi2(3)        =      8368.10
                                                Prob > chi2         =      0.0000
```

## Two-step results

crime	Coef.	WC-Robust Std. Err.	z	P> z	[95% Conf. Interval]	
crime						
L1.	.6206864	.010649	58.29	0.000	.5998148	.641558
policepc	-1.099586	.0440954	-24.94	0.000	-1.186012	-1.013161
legalwage	-1.007016	.036046	-27.94	0.000	-1.077665	-.9363673
_cons	42.62879	.559864	76.14	0.000	41.53148	43.7261

## Instruments for differenced equation

GMM-type: L(2/.)crime L(1/.)policepc

Standard: D.legalwage

## Instruments for level equation

Standard: \_cons

# Arellano-Bover/Blundell-Bond estimator

- The Arellano-Bond estimator formed moment conditions using lagged-levels of the dependent variable and the predetermined variables with first-differences of the disturbances
- [Arellano and Bover(1995)] and [Blundell and Bond(1998)] found that if the autoregressive process is too persistent, then the lagged-levels are weak instruments
- These authors proposed using additional moment conditions in which lagged differences of the dependent variable are orthogonal to levels of the disturbances
  - To get these additional moment conditions, they assumed that panel-level effect is unrelated to the first observable first-difference of the dependent variable

# xtdpdsys

- Use `xtdpdsys` to estimate parameters using the Arellano-Bover/Blundell-Bond estimator
- has syntax similar to `xtabond`

## xtdpdsys II

```
. xtdpdsys crime legalwage , twostep vce(robust) pre(policepc)
```

```
System dynamic panel-data estimation      Number of obs      =      7000
Group variable: id                       Number of groups   =      1000
Time variable: t

Obs per group:   min =      7
                 avg =      7
                 max =      7

Number of instruments =      63           Wald chi2(3)       = 11746.19
                                           Prob > chi2        =      0.0000
```

```
Two-step results
```

crime	Coef.	WC-Robust Std. Err.	z	P> z	[95% Conf. Interval]	
crime						
L1.	.6239032	.0089022	70.08	0.000	.6064553	.6413511
policepc	-1.07781	.0280265	-38.46	0.000	-1.132741	-1.022879
legalwage	-1.035179	.032075	-32.27	0.000	-1.098045	-.9723135
_cons	42.78394	.5334199	80.21	0.000	41.73845	43.82942

```
Instruments for differenced equation
```

```
  GMM-type: L(2/.)crime L(1/.)policepc
```

```
  Standard: D.legalwage
```

```
Instruments for level equation
```

```
  GMM-type: LD.crime D.policepc
```

```
  Standard: _cons
```

## xtdpd

- `xtabond` and `xtdpdsys` determine which instruments to create based on the assumption of no serial correlation the model you specify
- `xtabond` and `xtdpdsys` allow you to place limits on the number of lags used as instruments, but these commands are designed to do the work for you
- When you need to estimate the parameters of a model under weaker conditions, you need to create the instruments for your model yourself
- Use `xtdpd` for this case
  - `xtdpd` has a more complicated syntax that allows you more flexibility
- The next example uses `xtdpd` to produce the same estimates using `xtdpdsys`

## xtdpd II

```

. xtdpd L(0/1).crime legalwage policepc,          ///
>   dgmiv(crime ) dgmiv(policepc, lag(1 .))      ///
>   lgmiv(crime) lgmiv(policepc, lag(0))         ///
>   div(legalwage) twostep vce(robust)

Dynamic panel-data estimation          Number of obs      =       7000
Group variable: id                    Number of groups   =       1000
Time variable: t

Obs per group:  min =         7
                  avg =         7
                  max =         7

Number of instruments =        63          Wald chi2(3)       =    11746.19
                                           Prob > chi2        =       0.0000

```

## Two-step results

crime	Coef.	WC-Robust Std. Err.	z	P> z	[95% Conf. Interval]	
crime						
L1.	.6239032	.0089022	70.08	0.000	.6064553	.6413511
legalwage	-1.035179	.032075	-32.27	0.000	-1.098045	-.9723135
policepc	-1.07781	.0280265	-38.46	0.000	-1.132741	-1.022879
_cons	42.78394	.5334199	80.21	0.000	41.73845	43.82942

```

Instruments for differenced equation
  GMM-type: L(2/.)crime L(1/.)policepc
  Standard: D.legalwage
Instruments for level equation
  GMM-type: LD.crime D.policepc
  Standard: _cons

```

# The benefits of flexibility

- The flexibility of `xtdpd` allows you to estimate the parameters of models that `xtabond` and `xtdpdsys` cannot estimate
  - Models with predetermined or endogenous variables that do not have a lagged dependent variable
  - Models containing moving-average serial correlation in the residuals

# Predetermined variable, no lagged dependent variable

```
. xtdep crime legalwage policepc,          ///
>      dgmiv(policepc, lag(1 .)) lgmiv(policepc, lag(0))  ///
>      div(legalwage) twostep vce(robust)

Dynamic panel-data estimation      Number of obs      =      8000
Group variable: id                Number of groups   =      1000
Time variable: t

Obs per group:   min =      8
                 avg =      8
                 max =      8

Number of instruments =      37      Wald chi2(2)      =      494.14
                                   Prob > chi2          =      0.0000
```

## Two-step results

crime	Coef.	WC-Robust Std. Err.	z	P> z	[95% Conf. Interval]	
legalwage	-1.936197	.0905709	-21.38	0.000	-2.113713	-1.758681
policepc	.8066107	.08572	9.41	0.000	.6386026	.9746188
_cons	36.07915	1.547977	23.31	0.000	33.04517	39.11313

## Instruments for differenced equation

GMM-type: L(1/.)policepc

Standard: D.legalwage

## Instruments for level equation

GMM-type: D.policepc

Standard: \_cons

# Instruments for a model with MA(1) errors II

- Consider a model with MA(1) errors

$$y_{it} = \alpha y_{it-1} + \beta x_{it} + \nu_i + \epsilon_{it} + \gamma \epsilon_{it-1}$$

where the  $\epsilon_{it}$  are assumed to be IID and  $x_{it}$  is assumed to be strictly exogenous.

- Because the composite error,  $\epsilon_{it} + \gamma \epsilon_{it-1}$ , is MA(1), only lags two or higher of  $\Delta y_{it}$  are valid instruments for the level equation, assuming the initial condition that  $E[\nu_i \Delta n_{i2}] = 0$ .
  - Lagging the above equation two periods shows that  $\epsilon_{it-2}$  and  $\epsilon_{it-3}$  appear in the equation for  $y_{it-2}$
  - Because the  $\epsilon_{it}$  are IID,  $\Delta y_{it-2}$  is a valid instrument for the level equation with errors  $\nu_i + \epsilon_{it} + \gamma \epsilon_{it-1}$   
 ( $y_{it-2}$  will be correlated with  $y_{it-1}$  but uncorrelated with the errors  $\nu_i + \epsilon_{it} + \gamma \epsilon_{it-1}$ .  
 An analogous argument works for higher lags.

# Instruments for a model with MA(1) errors II

- First-differencing the model equation yields

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \beta \Delta x_{it} + \Delta \epsilon_{it} + \gamma \Delta \epsilon_{it-1}$$

- Because  $\epsilon_{it-2}$  is the farthest lag of  $\epsilon_{it}$  that appears in the differenced equation, lags three or higher are valid instruments for the differenced composite errors
  - Lagging the level equation three periods shows that only  $\epsilon_{it-3}$  and  $\epsilon_{it-4}$  appear in the equation for  $y_{it-3}$
  - So  $n_{it-3}$  is a valid instrument for the current differenced equation
  - An analogous argument works for higher lags.

## xtdpd V

```

. xtdpd L(0/1).crime legalwage policepc,          ///
>      dgmiv(crime, lag(3) ) dgmiv(policepc, lag(2 .))  ///
>      lgmmiv(crime) lgmmiv(policepc, lag(1)) div(legalwage)
Dynamic panel-data estimation          Number of obs      =      7000
Group variable: id                    Number of groups   =      1000
Time variable: t

Obs per group:  min =      7
                  avg =      7
                  max =      7

Number of instruments =      50          Wald chi2(3)      = 21227.62
                                           Prob > chi2       = 0.0000

```

## One-step results

crime	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
crime						
L1.	.6593019	.0079458	82.97	0.000	.6437283	.6748754
legalwage	-1.053813	.0348676	-30.22	0.000	-1.122152	-.9854736
policepc	-1.064979	.0291366	-36.55	0.000	-1.122086	-1.007873
_cons	42.79815	.5834091	73.36	0.000	41.65469	43.94161

## Instruments for differenced equation

GMM-type: L(3/.)crime L(2/.)policepc

Standard: D.legalwage

## Instruments for level equation





GMM-type: LD.crime LD.policepc

Standard: \_cons

- The Sargan statistic no longer rejects the model

```
. estat sargan
Sargan test of overidentifying restrictions
H0: overidentifying restrictions are valid
chi2(46)      = 74.18845
Prob > chi2   = 0.0053
```

## Bibliography

-  Anderson, T., and C. Hsiao. 1981.  
Estimation of Dynamic Models with Error components.  
*Journal of the American Statistical Association* 76(375): 598–606.
-  Arellano, M., and S. Bond. 1991.  
Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations.  
*Review of Economic Studies* 58: 277–297.
-  Arellano, M., and O. Bover. 1995.  
Another look at instrumental variables estimation of error-component models.  
*Journal of Econometrics* 68: 29–51.
-  Blundell, R., and S. Bond. 1998.  
Initial conditions and moment restrictions in dynamic panel-data models.  
*Journal of Econometrics* 87: 115–143.



Holtz-Eakin, D., W. Newey, and H. S. Rosen. 1988.  
Estimating Vector Autoregressions with Panel data.  
*Econometrica* 56(6): 1371–1395.



Windmeijer, F. 2005.  
A finite sample correction for the variance of linear efficient two-step  
GMM estimators.  
*Journal of Econometrics* 126(1): 25–51.

## xtdpd III

- `dgmmiv(varlist [ , lagrange(flag [llag]) ])` specifies GMM-type instruments for the differenced equation. Levels of the variables are used to form GMM-type instruments for the difference equation.
- `lgmmiv(varlist [ , lag(#)])` specifies GMM-type instruments for the level equation. Differences of the variables are used to form GMM-type instruments for the level equation.
- `iv(varlist [ , nodifference])` specifies standard instruments for both the differenced and level equations.
- `div(varlist [ , nodifference])` specifies additional standard instruments for the differenced equation.
- `liv(varlist)` specifies additional standard instruments for the level equation.

## xtdpd IV

- `dgmmiv(varlist [ , lagrange(flag [llag]) ])`  
All possible lags are used, unless `lagrange(flag llag)` restricts the lags to begin with `flag` and end with `llag`.
- `lgmmiv(varlist [ , lag(#)])`  
The first lag of the differences is used unless `lag(#)` specifies that `#`th lag of the differences should be used.
- `iv(varlist [ , nodifference])`  
Differences of the variables are used as instruments for the differenced equations, unless `nodifference` is specified, which requests that levels are to be used. Levels of the variables are used as instruments for the level equations.
- `div(varlist [ , nodifference])`  
Differences of the variables are used, unless `nodifference` is specified which requests that levels of the variables are to be used as instruments for the differenced equation.
- `liv(varlist)`  
Levels of the variables are used as instruments for the level equation.