Linear Dynamic Panel-Data Estimation using Maximum Likelihood and Structural Equation Modeling

Richard Williams, University of Notre Dame (rwilliam@nd.edu)
Paul Allison, University of Pennsylvania (allison@statisticalhorizons.com)
Enrique Moral-Benito, Banco de Espana, Madrid (enrique.moral@gmail.com)

Stata Conference, Columbus, Ohio
July 30, 2015

- Panel data (also sometimes known as longitudinal data or cross-sectional time series data, where data on the same subjects is collected at multiple points in time) have two big attractions for making causal inferences
  - The ability to control for unobserved, time-invariant confounders
  - The ability to determine the direction of causal relationships
- Controlling for unobservables can be accomplished with fixed effects methods that are well known
- For examining causal direction, the most popular approach has long been the cross-lagged panel model.
  - In cross-lagged panel models, *x* and *y* at time *t* affect both *x* and *y* at time *t*+1.

- Unfortunately, attempting to combine fixed effects models with cross-lagged panel models leads to serious estimation problems*
  - Economists typically refer to such models as *dynamic panel models* because of the lagged effect of the dependent variable on itself.
  - The estimation difficulties include error terms that are correlated with predictors, the so-called “incidental parameters problem”, and uncertainties about the treatment of initial conditions
  - * For reviews of the extensive literature on dynamic panel data models, see Wooldridge (2010), Baltagi (2013), or Hsiao (2014)

- The most popular econometric method for estimating dynamic panel models is the generalized method of moments (GMM) that relies on lagged variables as instruments.

- This method has been incorporated into several commercial software packages, usually under the name of Arellano-Bond (A-B) estimators.
  - For example, Stata has the `xtabond` and `xtabond2` commands

- While the A-B approach provides consistent estimators of the coefficients, there is substantial evidence that the estimators are not fully efficient (Ahn and Schmidt 1995) and often perform poorly when the autoregressive parameter (the effect of a variable on itself at a later point in time) is near 1.0.

- Moral-Benito (2013; see also Bai 2013) shows that Maximum Likelihood Estimation can be accomplished in a way that eliminates the incidental parameters problem and any need for special assumptions about initial conditions.

- Moral-Benito uses two equations to write his model. They are

(1) $y_{it} = \rho y_{it-1} + x'_{it}\beta + \omega_i'\delta + \alpha_i + \xi_t + \upsilon_{it}$

where

y_{it-1} is a vector of the lagged values of y

x_{it} is a vector of sequentially exogenous/predetermined time-varying variables

ω_i is a vector of time-invariant strictly exogenous variables

α_i is the unobservable time-invariant fixed effect

ξ_t captures unobserved common factors across units in the panel

υ_{it} is the time-varying error term

$$(2) \quad E(\upsilon_{it} \mid y_i^{t-1}, x_i^{t,'}, \omega_i, \alpha_i) = 0 \;\left(t = 1, \ldots, T\right)\left(i = 1, \ldots, N\right)$$

where

x_i^t denotes a vector of the observations accumulated up to t. This implies, for example, that the residual for y_5 is uncorrelated with predetermined variable x at times 1-5, but could be correlated with x at later times, e.g. x_6, x_7, etc. Put another way, predetermined variable x could be affected by earlier values of the dependent variable.

Other notation is as before

Condition (2) is the only assumption required for consistency and asymptotic normality (under fixed T when N tends to infinity)

Moral-Benito’s (2013) model does NOT include strictly exogenous time-varying variables but it can be easily modified to do so.

The meaning of each type of variable will become clearer as we move along.

- Allison (2014; in progress) further shows that the dynamic panel model is a special case of the general linear structural equation model (SEM) and that the method of Moral-Benito can be implemented (and extended) with Stata’s `sem` command.

- Allison (2014) and Moral-Benito (2013) claim that the SEM approach has several advantages over both GMM methods and previous ML methods:
  - there is no "incidental parameters" problem
  - initial conditions are treated as completely exogenous and do not need to be modeled
  - no difficulties arise when the autoregressive parameter is at or near 1.0
  - missing data are easily handled by full-information maximum likelihood
  - coefficients can be estimated for time-invariant predictors. (The A-B method cannot do this because it uses difference scores which causes all time-invariant variables to drop out)
  - many model constraints can be easily relaxed and/or tested
  - It is well known that likelihood-based approaches (ML) are preferred to method-of-moments (GMM) counterparts in terms of finite-sample performance (see Anderson, Kunitomo, and Sawa 1982), and that ML is more efficient than GMM under normality. Moral-Benito (2013) compares the widely-used panel GMM estimator of Arellano-Bond (1991) with its likelihood-based counterpart and confirms these results in the case of dynamic panel models with predetermined regressors.

- However, coding the sem method is both tedious and error prone

- Hence we introduce a command named `xtdpdml` with syntax similar to other Stata commands for linear dynamic panel-data estimation.

- `xtdpdml` greatly simplifies the SEM model specification process

Example 1: sem command vs xtdpdml command

- Allison reanalyzes data described by Cornwell and Rupert (1988) for 595 household heads who reported a non-zero wage in each of 7 years from 1976 to 1982.

  - wks = number of weeks employed in each year
  - union = 1 if wage set by union contract, else 0, in each year
  - lwage = ln(wage) in each year
  - ed = years of education in 1976

SEM coding (Adapted from Allison 2014 Appendix B)

```
use http://www3.nd.edu/~rwilliam/statafiles/wages, clear
keep wks lwage union ed id t
reshape wide wks lwage union, i(id) j(t)
sem      (wks2 <- wks1@b1 lwage1@b2 union1@b3 ed@b4 Alpha@1 E2@1 ) ///
         (wks3 <- wks2@b1 lwage2@b2 union2@b3 ed@b4 Alpha@1 E3@1) ///
         (wks4 <- wks3@b1 lwage3@b2 union3@b3 ed@b4 Alpha@1 E4@1) ///
         (wks5 <- wks4@b1 lwage4@b2 union4@b3 ed@b4 Alpha@1 E5@1) ///
         (wks6 <- wks5@b1 lwage5@b2 union5@b3 ed@b4 Alpha@1 E6@1) ///
         (wks7 <- wks6@b1 lwage6@b2 union6@b3 ed@b4 Alpha@1), ///
         var(e.wks2@0 e.wks3@0 e.wks4@0 e.wks5@0 e.wks6@0) var(Alpha) ///
         cov(Alpha*(ed)@0) cov(Alpha*(E2 E3 E4 E5 E6)@0) ///
         cov(_OEx*(E2 E3 E4 E5 E6)@0) cov(E2*(E3 E4 E5 E6)@0) ///
         cov(E3*(E4 E5 E6)@0) cov(E4*(E5 E6)@0) cov(E5*(E6)@0) ///
         cov(union3*(E2)) cov(union4*(E2 E3)) cov(union5*(E2 E3 E4)) ///
         cov(union6*(E2 E3 E4 E5)) ///
         iterate(250) technique(nr 25 bhhh 25) noxconditional
```

Practical Problems with SEM Coding

- Data need to be in wide format; most dynamic panel data sets will be in long format
- Coding is lengthy and error prone; getting the covariance structure right is especially difficult
- Output is voluminous and highly repetitive because of all the equality constraints
- Limitations of Stata make the coding less straightforward than we might like
  - Stata won't allow covariances between predetermined Xs and the Y residuals. `xtdpdml` therefore zeroes out most of the Y residuals and replaces them with latent exogenous variables (E2, E3, etc.)
  - Stata sometimes falsely claims a model is not identified when it really is
  - Some alternative/equivalent codings result in convergence problems or even fatal errors

Equivalent coding using xtdpdml

```
. use http://www3.nd.edu/~rwilliam/statafiles/wages, clear
. xtset id t
. xtdpdml wks L.lwage, inv(ed) pre(L.union)

Highlights parameterization:
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
wks2         |
        wks1 |   .1871266   .0201939     9.27   0.000     .1475473    .2267059
      lwage1 |   .6417879   .4842305     1.33   0.185    -.3072865    1.590862
      union1 |   -1.19136   .5168948    -2.30   0.021    -2.204455   -.1782652
          ed |  -.1122268   .0559478    -2.01   0.045    -.2218824   -.0025712
------------------------------------------------------------------------------
Number of units = 595. Number of periods = 6.
LR test of model vs. saturated: chi2(71)  =     110.23, Prob > chi2 =  0.0020
Wald test of all coeff = 0: chi2(4) =      90.09, Prob > chi2 =  0.0000
```

- One short command generates the equivalent of the 13 lines of `sem` code shown earlier. `xtdpdml` also handled temporarily reshaping the data to wide format.
- By default, all variable effects (but not the constants) are constrained to be equal across time. Therefore only the first equation (in this case for time 2) needs to be presented
- The LR statistic provides an overall goodness of fit test.
- The Wald statistic tests whether the effects of any of the variables in the model significantly differ from zero

- That is obviously a much simpler syntax. The reason it isn't simpler still (and why the `sem` coding is so difficult) is because there are several types of independent variables in the model
  - The lag 1 value of y (e.g. L1.wks) is included by default.
    - This can be changed with the ylag option, e.g. ylag(1 2), ylag(2 4)
    - ylag(0) will cause no lagged values of y to be included
  - Strictly exogenous variables are those that (by assumption) are uncorrelated with the error terms at all points in time. Equivalently, we assume that they are not affected by prior values of the dependent variable.
    - These variables are specified on the left side of the comma
    - Time series notation can be used, e.g. `xtdpdml y L1.lwage L2.lwage` would include the first and second lagged values of wages as independent variables.

- Predetermined variables, also known as sequentially or weakly exogenous, are variables that can be affected by prior values of the dependent variables.
  - In the current example, we allow for the possibility that weeks worked in one year can affect union status in later years
  - Time series notation can be used.
  - Predetermined variables are specified with the pre option.
  - Mechanically, the Y residuals are allowed to correlate with the later-in-time values of the predetermined variables.

- Time-invariant variables are variables whose values are constant across time, such as year born.
  - In the current example, years of education does not vary across time
  - These are specified with the inv option
  - The ability to use time-invariant variables in the model is one of the key advantages of the sem approach.

- Also automatically included in each model is the latent exogenous variable Alpha.
  - Alpha reflects the fixed effects that are common to each equation across time.
  - Alpha can freely covary with all the time-varying observed exogeneous variables (but not with the time-invariant observed exogeneous variables). As Allison says, “This is exactly what we want to achieve in order for Alpha to truly behave as a set of fixed effects”
  - The effect of Alpha is fixed at 1 in each equation (unless the alphafree option is specified)

Example 2: xtdpdml vs xtabond (real data)

```
webuse abdata, clear
keep if year >=1978 & year <= 1982
xtabond n l(0/1).w l(0/2).(k ys) yr1976-yr1984, lags(2)
xtdpdml n l(0/1).w l(0/2).(k ys) , ylags(1 2) tfix
```

All cases have data for 1978-1982, making the panel that is analyzed strongly balanced.

Syntax for the two commands is fairly similar in this case.

Time dummies are added to `xtabond` because, by default, `xtdpdml` allows the constants to differ across time.

	xtabond	xtdpdml
L.n	0.864	0.937***
	(1.35)	(7.01)
L2.n	-0.269	-0.182
	(-1.44)	(-1.88)
w	-0.616***	-0.649***
	(-5.49)	(-6.89)
L.w	0.227	0.304*
	(0.81)	(2.54)
k	0.364***	0.324***
	(4.34)	(5.71)
L.k	-0.201	-0.174*
	(-0.72)	(-2.20)
L2.k	0.0300	0.00824
	(0.26)	(0.12)
ys	0.575**	0.575***
	(2.80)	(3.32)
L.ys	-0.690*	-0.789***
	(-2.03)	(-3.94)
L2.ys	-0.0621	-0.0893
	(-0.25)	(-0.41)

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

- At least in this relatively simple example, the coefficients are very similar

- `xtdpdml` produces smaller standard errors and bigger t values, which is consistent with our earlier points about the advantages of ML over GMM especially in finite samples

- `xtdpdml` can also do things that `xtabond` can't, like include time-invariant variables in the model

Example 3: xtdpdml vs xtabond (simulated data)

	Bias in lagged Y (true = .75)		Bias in predetermined X (true = .25)	
	xtabond	xtdpdml	xtabond	xtdpdml
N = 100	-.015005	.0037066	-.0082087	.0018838
N = 500	-.002968	.000475	-.0016843	.000128
N = 1000	-.001151	.0004413	-.0005852	.0002762

* In each case 1,000 simulations are run. Adapted from Moral-Benito (2013), Table 1, Design 5.

- At least in these simulations, `xtdpdml` produces estimates that are closer to the true values than does `xtabond` (`xtdpdml` standard errors also tend to be smaller)

- As we would expect, advantages of the `ML/xtdpdml` method are greatest when the sample size is small

- Several other simulations suggest that `xtdpdml` tends to do as well or better as other alternatives (although more conditions need to be tested)

Alternatives to xtdpdml

- The user-written routines `xtmoralb` (Moral-Benito 2013) and `xtdpdqml` (Kripfganz, 2015; available from SSC) can do some of the same things as `xtdpdml`, and may be very useful in many situations. However, they also have some important limitations.
  - `xtmoralb` works extremely well with predetermined variables (indeed we used it to refine `xtdpdml`). However, it cannot handle time-invariant variables, lagged exogenous variables, and is not fully efficient with strictly exogenous variables.
  - `xtdpdqml` works with strictly exogenous variables and can also sometimes produce results very similar to `xtdpdml`. However, it cannot handle time-invariant variables (in a fixed effects model) and (according to the author) is inappropriate for predetermined variables. Also, `xtdpdqml` implements the ml method of Hsiao et al (2002) which makes strong and questionable assumptions about initial conditions

Other useful features of xtdpdml

- Can relax/impose/test constraints, e.g. `xfree` relaxes the constraint that the effects of the exogenous variables are invariant across time
- `details` shows the complete sem output
- `showcmd` shows the `sem` command that was generated. You can copy and edit this if `xtdpdml` can't estimate the exact model you want.
- The `fiml` option causes Full Information Maximum Likelihood to be used for missing data; default is listwise deletion
- `semopts(options)` lets additional sem options be included in the generated sem command

- Many/most `sem` postestimation commands can be used. You may need to use the `staywide` option to get some options to work.
  - For example, you could use `estat summarize` or `estat mindices`.
  - These options can help to assess model fit and identify areas where the model could be improved, e.g. the modification indices might suggest that some variables specified as strictly exogenous should be specified as predetermined instead.

Areas needing further study and/or program development

- We want to make `xtdpdml` output look more like the output from programs like `xtabond`, e.g. use lag notation for variable names

- Procedure works very well with strongly balanced panels with complete data. We need to examine how well the procedure works with unbalanced panels and missing data
  - By default, `sem` deletes cases on a listwise basis. Because data are converted to wide format, a missing wave or even missing data on a single variable can cause all the data for all waves for a case to be lost.
  - The abdata provided with Stata has 140 cases with 8 waves of data; but if you try to analyze all 8 waves only 14 cases are left!
  - Use of fiml (Full Information Maximum Likelihood) may help a lot, especially if there is only a little missing data, but it probably has its limitations

Additional Information

- Until the final version of xtdpdml is released on SSC, the beta version of the program (still subject to major revisions and use at your own risk) is available by typing the following from within Stata:

```
net install xtdpdml, from(http://www3.nd.edu/~rwilliam/stata)
```

- For more information see

  http://www3.nd.edu/~rwilliam/dynamic/

References

- Ahn, S. C. and Peter Schmidt (1995) “Efficient Estimation of Models for Dynamic Panel Data.” Journal of Econometrics 68: 5-27.
- Allison, Paul. 2014. “Maximum Likelihood for Dynamic Panel Models with Cross-Lagged Effects”. http://statisticalhorizons.com/wp-content/uploads/ML-DynamicPanel-1SP.pdf
- Allison, Paul. 2015. "Don't Put Lagged Dependent Variables in Mixed Models.“ http://statisticalhorizons.com/lagged-dependent-variables
- Arellano, M. and S. Bond (1991) “Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations.” The Review of Economic Studies 58: 277-297.
- Bai, Jushan (2013). “Fixed effects dynamic panel data models, a factor analytical approach.” Econometrica 81 (1): 285-314.
- Baltagi, Badi H. (2013), Econometric Analysis of Panel Data. Fifth Edition. New York: John Wiley & Sons.
- Hsiao, Cheng (2014) Analysis of Panel Data. Third Edition. London: Cambridge University Press.
- Hsiao, C., M. H. Pesaran, and A. K. Tahmiscioglu. 2002. Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. Journal of Econometrics 109: 107-150.
- Kripfganz, S. 2015. xtdpdqml: Quasi-Maximum Likelihood Estimation of Linear Dynamic Panel Data Models in Stata. Manuscript. Goethe University Frankfurt. http://www.kripfganz.de
- Moral-Benito, Enrique. 2013. "Likelihood-based Estimation of Dynamic Panels with Predetermined Regressors." Journal of Business and Economic Statistics 31:4, 451-472.
- Williams, Richard, Paul Allison and Enrique Moral-Benito. 2015. "Linear Dynamic Panel-Data Estimation using Maximum Likelihood and Structural Equation Modeling". Presented July 30, 2015 at the 2015 Stata Users Conference in Columbus, Ohio.
- Wooldridge, Jeffrey M. (2010) Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press.