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

Richard Williams, University of Notre Dame (rwilliam@nd.edu)

Paul D. Allison, University of Pennsylvania (allison@statisticalhorizons.com)

Enrique Moral-Benito, Banco de Espana, Madrid (enrique.moral@gmail.com)

Last Revised May 6, 2018

The article this presentation is based on is

forthcoming in *The Stata Journal*

- 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

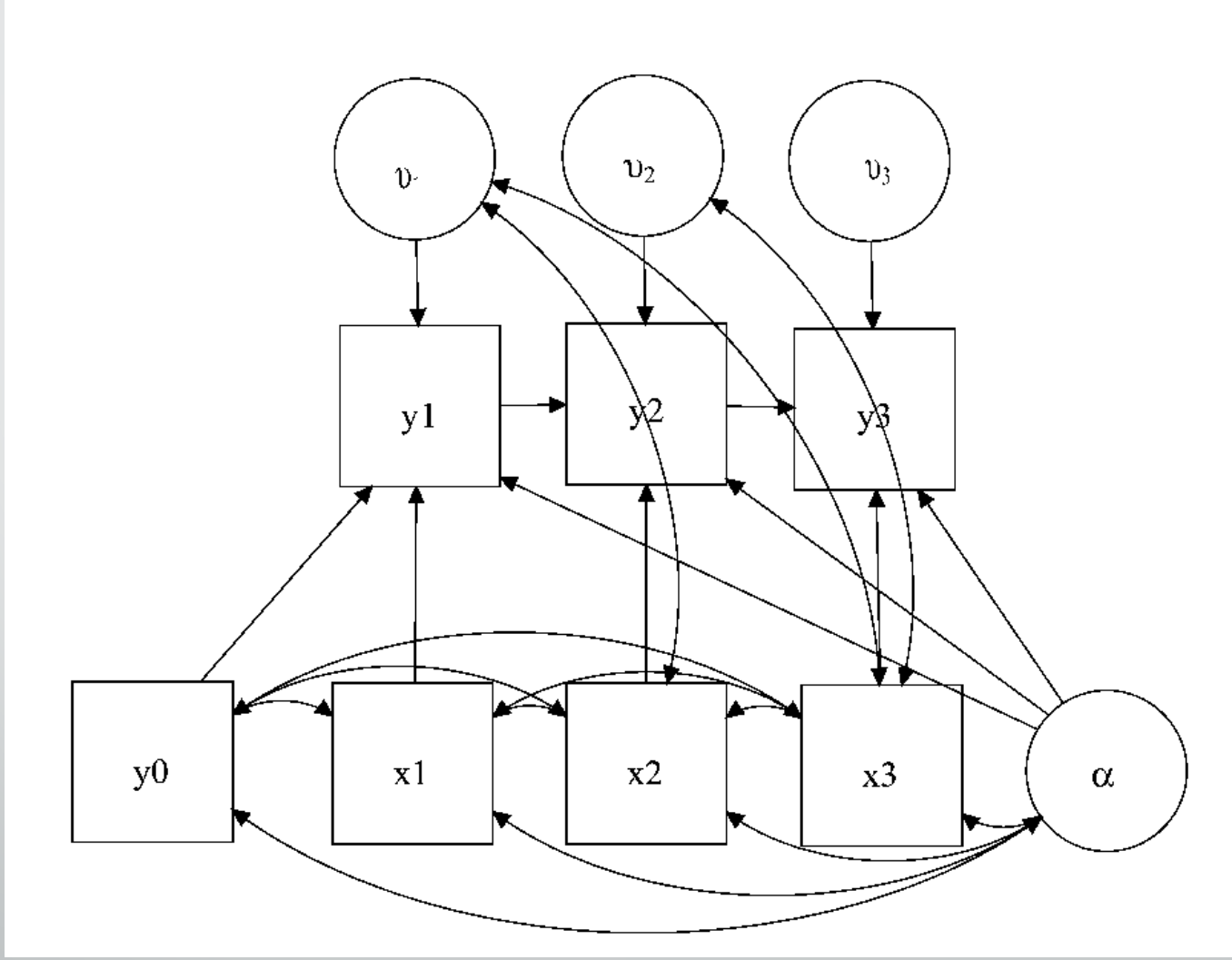
- The most popular econometric method for estimating dynamic panel models is the generalized method of moments (GMM) that relies on lagged variables as instruments.

- 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 AB approach provides consistent estimators of the coefficients, there is evidence that the estimators are not fully efficient, have considerable small-sample bias, 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.
- In our paper and related work, we show how Moral-Benito's models can be replicated and extended using SEM software widely available in software programs such as Mplus, Stata, and R
  - We also introduce a Stata program, xtdpdml, that greatly simplifies the model specification process.

Figure 1. Path Diagram for Dynamic Panel Model with *T*=3.



- Automatically included in each model is the latent exogenous variable Alpha.
  - Alpha reflects the fixed effects that are common to each equation across time. They are the effects of time-invariant variables not in the model.
  - Instead of relying on difference scores or other methods to eliminate the fixed effects, maximum likelihood estimation of this model is accomplished by allowing the fixed effects to have unrestricted correlations with the time-varying (but not time-invariant) predictors
  - This is exactly what we want to achieve in order for Alpha to truly behave as a set of fixed effects

- Cross-lagged causation can be accommodated by allowing the error term in each equation to correlate with future values of the time-dependent predictors. For example, x3 could be affected by y1 and y2. x would then be predetermined/ sequentially exogenous rather than strictly exogenous.
- By default, numerous parameters are fixed to be the same across waves, but these restrictions can be relaxed

- Using simulated data, Allison et al (2017) and Moral-Benito et al (forthcoming) show that the ML-SEM method outperforms the AB method with respect to bias and efficiency under most conditions. ML-SEM also has several other advantages over the AB method:
  - Time-invariant variables can be included in the model.
  - Missing values on predictors can easily be handled by full information maximum likelihood (FIML).
  - Error variances and other parameters can easily be allowed to vary with time.
  - Many goodness-of-fit measures are available to assess the over-identifying restrictions of the model.
  - There is no need to choose among many possible instrumental variables.

- 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

sem command vs xtdpdml command

- Allison et al (2017) reanalyze 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

```
use https://www3.nd.edu/~rwilliam/statafiles/wages, clear
keep wks lwage union ed id t
xtset 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 sometimes falsely claims a model is not identified when it really is
  - Some seemingly alternative/equivalent codings result in convergence problems or even fatal errors
  - Therefore you often have to use klutzy coding to make the model work in Stata

Equivalent coding using xtdpdml

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

Highlights: Dynamic Panel Data Model using ML for outcome variable wks
------------------------------------------------------------------------------
             |                 OIM
         wks |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
wks          |
         wks |
          L1. |   .1871266   .0201939     9.27   0.000     .1475473    .2267059
             |
       lwage |
          L1. |   .6417917   .4842304     1.33   0.185    -.3072823    1.590866
             |
       union |
          L1. |  -1.191349   .5168951    -2.30   0.021    -2.204445   -.1782536
             |
           ed |  -.1122267   .0559477    -2.01   0.045    -.2218822   -.0025711
------------------------------------------------------------------------------
# of units = 595. # of periods = 7. First dependent variable is from period 2.
Constants are free to vary across time periods
LR test of model vs. saturated: chi2(71)  =     110.23, Prob > chi2 =  0.0020
IC Measures: BIC =   25470.43  AIC =   24772.64
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 4.1 : ML/SEM vs GMM/Arellano-Bond (Adapted from Bollen and Brand 2010)

- The following examples are adapted from Bollen and Brand (2010).
  - They examine data from the National Longitudinal Survey of Youth. Respondents were 14 to 22 years old when first interviewed in 1979, and were interviewed annually or bi-annually for several years thereafter.
  - The dependent variable (lnwg) is log hourly wages in current job. The main independent variable (hchild) is total number of children the respondent had at the time of the interview.
  - Other variables in the model include whether or not married (mar) or divorced (div); educational attainment (eduatt); currently in school (cursc); several measures of part-time and full-time work experience (snrpt, snrft, exppt and expft); breaks in employment history (break); and time-invariant race and ethnicity measures (black, hisp).
  - The data set is strongly balanced, but several cases and records have missing data on one or more variables.

```
*** Section 4.1 -- Comparisons with AB, real data, using fiml and listwise
use https://www3.nd.edu/~rwilliam/statafiles/bollenbrand, clear
set matsize 7500
* Arellano-Bond
xtabond lnwg hchild marr div eduatt cursc snrpt snrft exppt expft break black hisp
estimates store gmm
* FIML
xtdpdml lnwg hchild marr div eduatt cursc snrpt snrft exppt expft break , ///
    constinv errorinv fiml tfix store(fiml) ///
    inv(black hisp) ti(Adapted from Bollen & Brand Social Forces 2010)
* Listwise deletion used instead of fiml
xtdpdml lnwg hchild marr div eduatt cursc snrpt snrft exppt expft break , ///
    constinv errorinv tfix store(normal) ///
    inv(black hisp) gof
```

Comparison of A/B & SEM approaches - Adapted from Bollen and Brand 2010

	(1) gmm	(2) fiml	(3) listwise
main			
L.lnwg	-0.00728	0.338***	0.278***
hchild	-0.00913	-0.0210***	-0.0145
marr	0.0468**	0.0360**	0.0591**
div	0.0747***	0.0617***	0.0578*
eduatt	0.0576***	0.0583***	0.0764***
cursc	-0.0811***	-0.108***	-0.0923***
snrpt	0.0133*	0.00885*	0.0151*
snrft	0.0141***	0.0174***	0.0102***
exppt	0.0566***	0.0309***	0.0410***
expft	0.0608***	0.0307***	0.0380***
break	0.0201**	0.0371***	0.0196**
black	0	-0.00746	-0.0300
hisp	0	0.0731***	0.0738***
_cons	0.628***		
N_g	3488	5285	1229

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

- The results are strikingly different.
  - Almost 21,000 records have data on at least one variable in the model, and all of these observations are used by xtdpdml (with the fiml option).
  - However, only 8,915 records are used by xtabond because it deletes any record with missing data. That is, GMM only uses about 42% of the data.
  - Perhaps for this reason, xtabond produces a highly implausible estimate of almost zero effect of lagged wages on current wages and also says that the effect of the main independent variable, number of children, is statistically insignificant.
  - In the xtdpdml results both effects are highly significant and the signs of the effects are in the expected direction.
  - Many other variables have larger z-statistics in xtdpdml than they do in xtabond.
  - xtabond cannot estimate effects for the time-invariant variables black and hisp, while xtdpdml shows that the effect of hisp is highly significant.

- Even if we leave out the fiml option, thereby deleting all persons who have missing data at any time point, the results from SEM/xtdpdml seem somewhat more plausible than those from GMM.
- Granted, we don't know what the true values of the parameters are.
- But, Monte Carlo simulations in our other papers show that the ML-SEM method is less biased and more efficient than the GMM method under a wide range of conditions.

Other Examples (see paper)

- 4.2. Panel Model wth fixed effects; Goodness of fit measures
  - This example shows how ML-SEM goodness of fit measures can be used to identify ways to improve model specification. For example, various equality constraints that are imposed by default can be relaxed.

- 4.3. Fixed Effects vs Random Effects: An Alternative to the Hausman test
  - The example shows how both FE and RE models can be estimated. With other approaches, a Hausman test can be used to compare the two. Hausman tests often have problems though. With ML-SEM, a likelihood ratio test can be used instead.

- 4.4. Non-Normality
  - Assumptions of multivariate normality will sometimes be violated. Often this is not a problem. When it is, the example shows how vce(sbentler), vce(robust), and method(adf) can be used to deal with non-normality.

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. `semfile` will output the generated code. 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

- Time-series notation can be used
- Interaction effects can be specified, but the procedure is sometimes different than it is for other techniques
- There are various options that may help with speed or convergence issues

- 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.

- The mplus option generates code that can estimate the models using Mplus. Mplus is usually much faster than Stata.
  - Hand-coding in Mplus would be even more tedious than hand-coding in Stata
- The lavaan option generates code that can estimate models using R's add-on lavaan package. R is free and lavaan seems to run about twice as fast as Stata
  - Hand-coding in R would be a total nightmare!!!

Limitations of xtdpdml

- Much slower than GMM routines like xtabond.
- Works best in large N / small T situations. Often will not work when T > 10.
- Interactions with time can be done, but it is done differently, e.g. use the xfree option instead of creating interactions.
- The paper and the help file offer more guidance for dealing with problems.

Additional Information

- The paper these slides are based on is forthcoming in *The Stata Journal.* A working paper version is at

https://www3.nd.edu/~rwilliam/dynamic/SJPaper.pdf

- Additional working papers and technical support materials are available at

https://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. 2015. "Don't Put Lagged Dependent Variables in Mixed Models." http://statisticalhorizons.com/lagged-dependent-variables
- Allison, Paul D., Richard Williams and Enrique Moral-Benito. 2017. "Maximum Likelihood for Cross-Lagged Panel Models with Fixed Effects." Socius 3: 1-17. http://journals.sagepub.com/doi/suppl/10.1177/2378023117710578
- 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.
- Bollen, Kenneth, and Jennie Brand. 2010. "A General Panel Model with Random and Fixed Effects: A Structural Equations Approach." Social Forces 89:1, 1-34.
- 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.
- Moral-Benito, Enrique, Paul D. Allison, and Richard Williams. Forthcoming in Applied Economics. "Dynamic Panel Data Modeling using Maximum Likelihood: An Alternative to Arellano-Bond." https://www3.nd.edu/~rwilliam/dynamic/Benito_Allison_Williams.pdf
- Williams, Richard, Paul D. Allison and Enrique Moral-Benito. Forthcoming in The Stata Journal. "xtdpdml: Linear Dynamic Panel-Data Estimation using Maximum Likelihood and Structural Equation Modeling." https://www3.nd.edu/~rwilliam/dynamic/SJPaper.pdf
- Wooldridge, Jeffrey M. (2010) Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press