Dynamic Panel Data Models with Irregular Spacing

SMU CORE Colloquium Series

Daniel L. Millimet[1] Ian McDonough[2]

[1]Southern Methodist University & IZA

[2]Southern Methodist University

February 2014

The Situation

- Many interesting outcomes or behaviors of interest in social and physical sciences are *dynamic* in nature
  - **Education:** Cumulative process whereby current academic achievement depends on current and past accumulation of knowledge
  - **Health:** Cumulative process whereby current health state depends on current inputs and past health
  - **Demography:** Current population depends on past population and changes
  - **Policy:** Partial adjustment implies current policy choice is a weighted average of past policy choice and desired adjustment
  - **Finance:** Asset returns have significant persistence
  - **Physical sciences:** Biological growth models, ???
- Analysis of dynamic processes in practice require either *time series* data or *panel* data
  - Time series: y_t, $t = 1, ..., T$
  - Panel: y_{it}, $i = 1, ..., N$; $t = 1, ..., T$

- Dynamic models have a long history in applied time series
  - Univariate models... Example: $AR(1)$

$$y_t = \rho y_{t-1} + \varepsilon_t$$

    - ★ Unit root testing is focused on $\mathrm{H}_o : \rho = 1$ vs. $\mathrm{H}_a : \rho < 1$
  - Multivariate models... Example: $VAR(p)$

$$\mathbf{Y}_t = \Pi \mathbf{Y}_{t-1} + \boldsymbol{\varepsilon}_t$$

    where $\mathbf{Y}$ is $p \times 1$, Π is $p \times p$, and $\boldsymbol{\varepsilon}$ is $p \times 1$
- Dynamic panel data (DPD) originated with Balestra and Nerlove (1966)

$$y_{it} = \gamma y_{it-1} + x_{it}\beta + \alpha_i + \varepsilon_{it}$$

  where α_i may be modeled as a *random* or *fixed* effect

The Problem

- Given access to *typical* time series data, estimation of univariate and multivariate dynamic models, testing for unit roots, etc. is (reasonably) straightforward
- Given access to *typical* panel data, estimation of DPD models is (reasonably) straightforward
  - First-differenced IV (Anderson & Hsiao 1981)
  - GMM (Arellano & Bond 1991)
  - System GMM (Arellano & Bover 1995; Blundell & Bond 1998)
  - Long-Differenced IV (Hahn et al. 2007)
  - Bias-Corrected LSDV (Kiviet 1995; Bun & Carree 2005)
  - Orthogonal to Backward Mean Transformation (Everaert 2012)

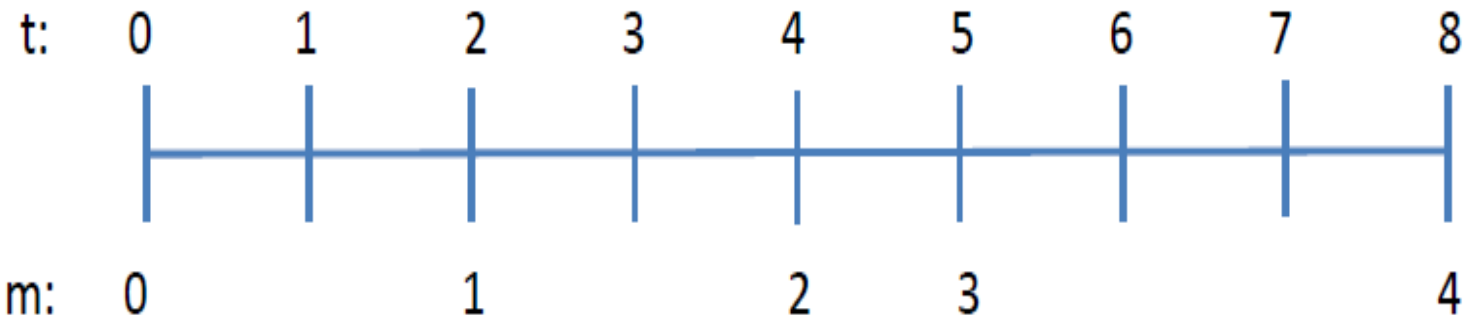
- Problem arises in all these contexts when *typical* data are unavailable
- In particular, we are concerned with the problem of *irregularly* spaced data

Definition

A data set is said to be irregularly spaced if successive periods of observed data do not conform to successive periods *as defined by the underlying data-generating process* (DGP).

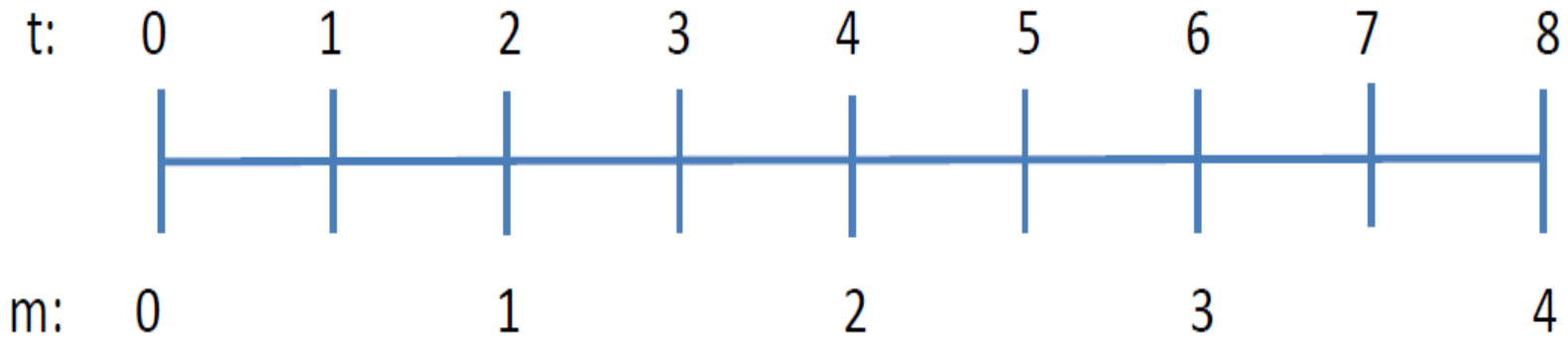
Hypothetical Example...

- Actual periods: $t = 0, 1, ..., 8$
- Observed periods: $t = 0, 2, 4, 5, 8$
- Denote the observed periods as $m = 0, 1, 2, 3, 4$
- Let $t(m)$ represent a mapping from the mth observed period to the corresponding t^{th} actual period



A bit of philosophy:

1. It is possible to *assume the problem away*
   - In time series, the distance between successive waves is referred to as the *observation interval*, whereas the *unit period* denotes the reference unit of time for the underlying process (Fuleky 2011)
   - In discrete time models in time series, it is commonplace to set the unit period equal to the observation interval (Hamilton 1994)
   - Whether this is convincing depends on the data structure and problem
2. Data collected at *uniform* intervals but *with gaps* are irregularly spaced according to our definition *if* the underlying DGP does not have such gaps; i.e., the unit period is smaller than the observation interval.



Comment

- The *same* periods are assumed to be observed for *all i*
  - Not the typical missing data scenario (not observation-specific missing)
  - Missingness is clearly 'at random'

Actual Examples...

- Time series
  - Stock market closings on holidays, weekends
  - Changes in historical data collection frequency (mixed frequency)
- Longitudinal surveys
  - McKenzie (2001) provides developing country examples
  - Table 1 in our paper lists developed country examples

Table 1. Some examples of unequally spaced surveys.

Panel A: household surveys for monitoring poverty in developing countries[a]		
Country	*Survey*	*Survey periods*
Bolivia	Encuesta Integrada de Hogares (EIH)	Mar 89, Nov 89, Sept 90, Nov 91, Nov 92, July–Dec 93, July–Dec 94, June 95
Brazil	Pesquisa Nacional por Amostra de Domicillios (PNAD)	Annual surveys since 1971, but surveys not taken in census years 1980 and 1991
Chile	Caracterización Socioeconómica Nacional (CASEN)	1985, 87, 90, 92, 94, 96
Ethiopia	Welfare monitoring survey	1995, 97, 98
Ghana	Ghana living standards survey	1987, 88, 91, 98
Kenya	Welfare monitoring survey	1992, 94, 97
Kyrgyz Republic	Poverty monitoring survey	1993, 96, 96, 97, 98
Mexico	Encuesta nacional de Ingreso-Gasto de los hogares (ENIGH)	1984, 89, 92, 94, 96
Nigeria	National consumer survey	1980, 85, 92, 96
Panama	Encuesta de Hogares–Mano de Obra (EMO)	1979, 89, 91, 95, 96
Peru	Encuesta Nacional de Hogares Sobre Medición de Niveles de Vida (ENNIV)	1985, 90, 91, 94
Senegal	Enquête Démographique et de Santé	1986, 92, 97
Thailand	Thailand Socio-Economic Survey (SES)	1975, 81, 86, 88, 90, 92, 94, 96, 98

Table 1. Examples of Irregularly Spaced Longitudinal Surveys in Developed Countries

Country	Survey	Structure
Australia	Australian Longitudinal Study on Women's Health (ALSWH)	Waves for different cohorts are separated by 2-4 years
	Longitudinal Study of Australian Children (LSAC)	Biennial from 2003-present
	Longitudinal Survey of Immigrants to Australia	Wave 1 covers first 6 months post-immigration, Wave 2 covers 6-18 months post-immigration, Wave 3 covers 18-42 months post-immigration
Canada	National Longitudinal Survey of Children and Youth (NLSCY)	Biennial from 1994-2009
France	French Longitudinal Study of Children (ELFE)	Individuals at birth, 2 months old, 10 months old, annual until 5 years old, then every 2-3 years thereafter
Japan	Nihon University Japanese Longitudinal Study of Aging (NUJLSOA)	Biennial for first three waves, Waves 3 and 4 separated by 3 years
UK	1958 National Child Development Study (NCDS)	Individuals at birth, 7 years old, 11 years old, 16 years old, 23 years old, 33 years old, 42 years old, 46 years old, 50 years old, and 55 years old
	1970 British Cohort Study (BCS70)	Individuals at birth, 5 years old, 10 years old, 16 years old, 26 years old, 30 years old, 34 years old, 38 years old, and 42 years old
	Millennium Cohort Study (MCS)	Individuals at birth, 9 months old, 3 years old, 5 years old, and 7 years old
	National Pupil Database (NPD)	Annual data on all students attending state primary and secondary schools, but national achievement assessments are only for ages 7, 11, 14, and 16

USA	Current Populaton Survey	4 consecutive months, 8 month gap, 4 consecutive months
	Early Childhood Longitudinal Survey-Kindergarten Cohort (ECLS-K)	Fall & Spring Kindergarten, Fall & Spring 1st Grade, Spring 3rd Grade, Spring 5th Grade, Spring 8th Grade
	Early Childhood Longitudinal Survey-Birth Cohort (ECLS-B)	Children at 9 months of age, 2 years old, 4 years old, Fall Kindergarten (either age 5 or 6)
	Education Longitudinal Study of 2002 (ELS:2002)	Tenth Grade, Twelfth Grade, 4 years post baseline, 10 years post baseline
	General Social Survey (GSS)	Biennial
	Health and Retirement Study (HRS)	Biennial
	High School & Beyond (HS&B)	2 cohorts surveyed every two years for 6 years with the younger cohort also surveyed 12 years post baseline
	High School Longitudinal Study of 2009 (HSLS:09)	Fall 9th Grade, Spring 11th Grade, Spring 12th Grade, 6 years post baseline, 12 years post baseline
	National Education Longitudinal Study of 1988 (NELS:88)	8th Grade, 10th Grade, 12th Grade, 6 years post baseline, 12 years post baseline
	National Longitudinal Study of Adolescent Health (Add Health)	Waves 1 and 2 are separated by 1 year, Wave 3 is 6 years post baseline, Wave 4 is 13 years post baseline
	National Longitudinal Survey of Youth 1979 (NLSY79)	Annual from 1979-1994, biennial thereafter
	Panel Study of Income Dynamics (PSID)	Annual from 1968-1997, biennial thereafter
	Second Longitudinal Study of Aging (LSOAII)	Waves 1 and 2 are separated by 3 years, Waves 2 and 3 are separated by 2 years
	Survey of Income and Program Participation	Variable design, but 8 equally-spaced waves are administered over a 32 month window

The Literature

- Time Series
  - **Univariate & multivariate models:** Savin and White (1978); Jones (1980, 1985, 1986); Dunsmuir and Robinson (1981); Harvey and Pierse (1984); Palm and Nijman (1984); Robinson (1985); Dufour and Dagenais (1985); Kohn and Ansley (1986); Shively (1993); Chiu et al. (2012)
  - **Unit root testing:** Shin and Sarkar (1994a,b); Ryan & Giles (1998)
- Panel
  - **Static models with AR(1) errors:** Jones and Boadi-Boateng (1991); Baltagi & Wu (1999)

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it}, \qquad \varepsilon_{it} = \rho\varepsilon_{it-1} + u_{it}$$

  - **DPD**
    - ★ **No unobserved effects:** Rosner & Munoz (1988)

$$y_{it} = \gamma y_{it-1} + x_{it}\beta + \varepsilon_{it}, \qquad \varepsilon_{it} \sim iid$$

    - ★ **Pseudo-panels:** McKenzie (2001)

The Details

- DGP given by

$$y_{it} = \gamma y_{it-1} + x_{it}\beta + \alpha_i + \varepsilon_{it}, \qquad i = 1, ..., N; \; t = 1, ..., T,$$

where

  - y_{it} = outcome for observation i in period t
  - γ is the autoregressive parameter ($|\gamma| < 1$)
  - x is a vector of covariates with associated parameter vector β
  - α_i is the unobserved effect
  - ε_{it} is the idiosyncratic, mean zero error term
  - y_{i0} is the initial condition

- Given a random sample, $\{y_{it}, x_{it}\}_{i=1,..,N;t=0,1,...,T}$, estimation is straightforward
  - Anderson & Hsiao (1981) suggest
    - ★ First-difference to eliminate α_i

$$\Delta y_{it} = \gamma \Delta y_{it-1} + \Delta x_{it}\beta + \Delta \varepsilon_{it}, \qquad i = 1, ..., N; \; t = 2, ..., T$$

    - ★ Instrument for Δy_{it-1} (x_{it-1} or y_{it-2} are valid IVs)
    - ★ Additional restrictions can be incorporated in a GMM framework (Arellano & Bond 1991; Arellano & Bover 1995; Blundell & Bond 1998)
  - Other transformations to eliminate α_i such as long-differencing
  - Bias-corrected fixed effects
  - Orthogonal to backward mean transformation

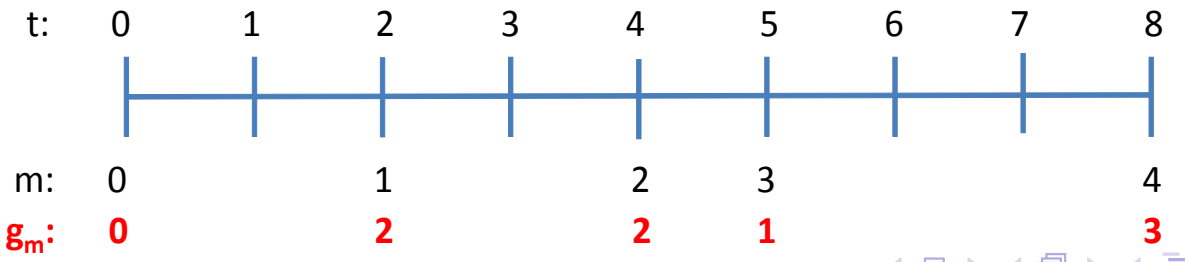
- Suppose we do not have a random sample, $\{y_{it}, x_{it}\}_{i=1,..,N;t=0,1,...,T}$
- Instead observe $\{y_{im}, x_{im}\}_{i=1,..,N;m=0,1,...,M}$

- With repeated substitution, the model defined over the *observed* periods is given by

$$y_{im} = \gamma^{g_m} y_{im-1} + x_{im}\beta + \underbrace{\left[\sum_{j=1}^{g_m-1} x_{i,t(m)-j}\gamma^j\beta + \left(\frac{1-\gamma^{g_m}}{1-\gamma}\right)\alpha_i + \sum_{j=0}^{g_m-1}\gamma^j\varepsilon_{i,t(m)-j}\right]}_{\text{Composite Error}}$$

where

- $i = 1, ..., N$; $m = 1, ..., M$
- g_m, $m = 1, ..., M$, is the *gap size* or the number of periods between observed period m and $m-1$
- $t(m)$ is the actual period reflected by observed period m
- Brackets contains all unobserved determinants of y_{im}



Three estimation difficulties arise:

1. The coefficient on the lagged dependent variable is not constant
2. The error term contains the covariates, x, and the idiosyncratic errors, ε, from the missing periods
3. The unobserved effect, α, now has a period-specific factor loading (as in the *interactive fixed effects* and *time-varying inefficiency* literatures)

Note: If $g_m = g > 1$ for all m, where g is any finite constant, then the data are *equally-spaced* but with *gaps* $\Rightarrow$ FD will eliminate α, but issues (1) and (2) remain.

$\Rightarrow$ All the standard DPD estimators are no longer valid!

- Re-write the model compactly as

$$\boxed{y_{im} = \gamma^{g_m} y_{im-1} + x_{im}\beta + \theta_m \alpha_i + \widetilde{\varepsilon}_{im}}$$

where

$$\theta_m \equiv \left(\frac{1-\gamma^{g_m}}{1-\gamma}\right); \qquad \widetilde{\varepsilon}_{im} \equiv \sum_{j=1}^{g_m-1} x_{i,t(m)-j}\gamma^j\beta + \sum_{j=0}^{g_m-1} \gamma^j \varepsilon_{i,t(m)-j}$$

Prior literature re-visited:

- Rosner & Munoz (1988)
  - Non-linear least squares (NLS)
  - Linear interpolation of x
  - No unobserved effects
- McKenzie (2001)
  - Non-linear least squares (NLS)
  - No covariates (or available from alternative data source)
  - Pseudo-panel $\Rightarrow$ averaging within cohort-time cells to eliminate α

We offer two new, potential solutions

1. Quasi-differencing
2. NLS-IV with and without imputation based on Everaert (2012)

Proposed Solution #1: Quasi-Differencing

- Quasi-differencing (QD) entails

$$y_{im} - \varphi_m y_{im-1} \quad \forall \varphi_m$$

- Letting

$$\varphi_m = \frac{\theta_m}{\theta_{m-1}} = \frac{1-\gamma^{g_m}}{1-\gamma^{g_{m-1}}}$$

simplifies to

$$\boxed{y_{im} - \varphi_m y_{im-1} = \gamma^{g_m} y_{im-1} - \varphi_m \gamma^{g_{m-1}} y_{im-2} + (x_{im} - \varphi_m x_{im-1})\beta + \tilde{\tilde{\varepsilon}}_{im}}$$

where

$$\tilde{\tilde{\varepsilon}}_{im} = f\left(x_{i,t(m-2)+1}, \ldots, x_{i,t(m-1)-1}, x_{i,t(m-1)+1}, \ldots, x_{i,t(m)-1}, \varepsilon_{i,t(m-2)+1}, \ldots, \varepsilon_{i,t(m)}\right)$$

- If φ_m were known, then a QD-NLS-IV estimator is feasible
  - x_{im-1} (and others) as IVs
  - $\widehat{\gamma}$, $\widehat{\beta}$ are consistent if x is strictly exogenous *and* serially uncorrelated
- With φ_m unknown, we propose a QD-GMM estimator which requires x to be strictly exogenous *and* serially uncorrelated

Monte Carlo

Design

- DGP

$$y_{it} = \gamma y_{it-1} + \beta x_{it} + \alpha_i + \varepsilon_{it}, \qquad i = 1, ..., N; \; t = -99, ..., 0, 1, .., T$$

- $T = 17$, $M = 6$, and observed periods include $\{0, 1, 2, 3, 7, 11, 17\}$
- $(\gamma, \beta) = (0.8, 0.8)$
- $N = 500$; Simulations $= 250$
- Two designs
  1. $\mathrm{Cov}(x_{it}, \alpha_i) \neq 0$, $\mathrm{Cov}(x_{it}, x_{it-1}) = 0$
  2. $\mathrm{Cov}(x_{it}, \alpha_i) \neq 0$, $\mathrm{Cov}(x_{it}, x_{it-1}) \neq 0$

Monte Carlo

Results

- x **strictly exogenous and serially uncorrelated**
  - QD-GMM performs well
  - Everaert-type estimator with the Mundlak approach also performs well
  - Traditional estimators (AH, AB, BB) perform reasonably well in some designs for β, but not γ
- x **strictly exogenous and serially correlated**
  - QD-GMM performs well
  - Everaert-type estimator with the Mundlak approach and $AR(1)$ imputation performs well
  - QD-GMM is generally better when γ is relatively low; Everaert otherwise

Simulation Results: $\gamma = 0.8$, $\beta = 0.8$.

Estimator	γ		β	
	DGP2	DGP4	DGP2	DGP4
AH	0.532	12.682	0.225	4.405
AB	0.069	0.074	0.117	0.264
BB	0.220	0.228	0.091	0.237
QD-GMM	0.035	0.030	0.081	0.102

Simulation Results: Long-run Coefficient.

Estimator	$\beta/(1-\gamma)$	
	DGP2	DGP4
AH	0.248	0.149
AB	0.228	0.118
BB	0.320	0.206
QD-GMM	0.080	0.082

Note: Numbers are RMSE.

Applications

Dynamics of Child Development

$$y_{it} = \gamma y_{it-1} + x_{it}\beta + \alpha_i + \varepsilon_{it}, \qquad i = 1, ..., N; \; t = 1, ..., T,$$

1. Value-added models for student test scores,
   where
   - y = student math or reading test score
   - x = current school inputs (e.g., class size, class behavior, teacher experiece, teacher certification status)
   - α = student family background, innate ability
2. Dynamic models for child obesity,
   where
   - y = child's BMI
   - x = current inputs (e.g., household characteristics, welfare/SNAP participation, school meals)
   - α = genetics

Applications

Data

- Early Childhood Longitudinal Survey – Kindergarten Cohort
  - Collected by US Dept of Education
  - Nationally representative sample of 20,000+ children entering kindergarten in Fall 1998
  - Dispersed across about 1,000 schools
  - Waves in Fall K, Spring K, Spring 1, Spring 3, Spring 5, and Spring 8
  - Sample sizes
    - ★ Achievement: $N = 5,977$, $NT = 29,855$ (math); $N = 5,564$, $NT = 27,820$
    - ★ BMI: $N = 9,155$, $NT = 54,930$

Applications

Results

- Estimator matters for degree of persistence in achievement
- Effects of other covariates are generally unaltered

Table 12. Determinants of Child Human Capital.

	AH	AB	BB	QD-GMM
I. Math z-Score				
γ	0.224	0.326	0.314	0.497
	(0.217)	(0.015)	(0.010)	(0.021)
II. Reading z-Score				
γ	0.775	0.571	0.520	0.709
	(0.559)	(0.016)	(0.011)	(0.017)
III. BMI z-Score				
γ	0.361	0.221	0.349	0.389
	(0.061)	(0.009)	(0.007)	(0.016)

Conclusion

- Traditional estimators of DPD models fair poorly when data are irregularly spaced
- QD-GMM performs well in many cases
- Does not perform particularly well when persistence is *high* and covariates are *serially correlated* and *really matter*
- Much work to be done given the popularity of irregularly spaced longitudinal surveys

Mind the gap!

Thank you!