Implementing the Keane and Runkle approach for fitting dynamic panel data models in Stata using XTKR

Michael Keane and Timothy Neal

University of New South Wales

CEPAR

August 20, 2019

Introduction

- Estimating dynamic panel data models that feature both fixed effects and lagged dependent variables is a nontrivial problem:
    - Standard FE-OLS and First-Difference estimators are inconsistent in N.
    - 2SLS can be consistent in N but is inefficient.
- Keane and Runkle (1992) proposed 'forward-filtering' to eliminate serial correlation and increase efficiency.
- Papers such as Arrelano and Bond (1991) and Blundell and Bond (1998) argued that this was not fully efficient as it did not use all available instruments.
- The estimators proposed in those papers, which were Dynamic GMM and System GMM respectively, relies on all available lags as instruments to further improve efficiency.

Introduction

- The GMM estimators can suffer from the “many-instruments problem”, which bias the coefficients towards OLS.
- Baltagi (2005) states that:

  “Ziliak (1997) performs an extensive set of Monte Carlo experiments for a dynamic panel-data model... [and] finds that the downward bias of generalized method of moments (GMM) is quite severe as the number of moment conditions expands, outweighing the gains in efficiency. Interestingly, Ziliak finds that the forward filter two-stage least-squares (2SLS) estimator proposed by Keane and Runkle (1992) performs best in terms of the bias/efficiency tradeoff and is recommended.”

Introduction

- In this presentation we will:
  - present a user-written Stata command called XTKR that implements the Keane and Runkle (1992) approach,
  - apply the estimator to an empirical application, and
  - present new Monte Carlo evidence that shows that this approach can perform better than the popular alternatives.
- In the Monte Carlo simulations we also consider methods proposed by Windmeijer (2005) and Roodman (2009) to compress the instrument set in the Arellano-Bond and Blundell-Bond approaches.

Dynamic Panel Data Models

- Consider the dynamic panel data model:

$$y_{it} = \beta_0 + \beta_1 y_{it-1} + \beta_2 \boldsymbol{x}_{it} + \mu_i + \epsilon_{it} \tag{1}$$

- OLS is biased and inconsistent because y_{it-1} is correlated with μ_i by construction.
- Applying the 'within' transformation (i.e. the Fixed Effects estimator) yields:

$$(y_{it} - \bar{y}_i) = \beta_1(y_{it-1} - \bar{y}_i) + \beta_2(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i) + (\epsilon_{it} - \bar{\epsilon}_i) \tag{2}$$

- But this is also biased because y_{it-1} and $\bar{y}_i$ are correlated with $\bar{\epsilon}_i$ by construction. This bias decreases as T increases, and is called the Nickell bias after Nickell (1981).

Dynamic Panel Data Models

- Another popular alternative is the First-Difference estimator:

$$(y_{it} - y_{it-1}) = \beta_1(y_{it-1} - y_{it-2}) + \beta_2(\boldsymbol{x}_{it} - \boldsymbol{x}_{it-1}) + (\epsilon_{it} - \epsilon_{it-1}) \quad (3)$$

- This is inconsistent because: (i) y_{it-1} is correlated with ϵ_{it-1} and also (ii) $\boldsymbol{x}_{it}$ is correlated with ϵ_{it-1} (assuming $\boldsymbol{x}_{it}$ is only weakly exogenous or predetermined).
- 2SLS applied to (1) or (3) can provide consistent estimates if appropriate instruments are used:
  - To estimate (1) using 2SLS, we can use Δy_{it-1} and $\Delta \boldsymbol{x}_{it-1}$ as instruments as they will be uncorrelated with ϵ_{it} and μ_{it}.
  - To estimate (3), we can use instruments that are uncorrelated with ϵ_{it-1}, such as y_{it-2} and $\boldsymbol{x}_{it-2}$.
- While 2SLS will be consistent, the presence of serially correlated errors will lead to inefficient estimates.

Dynamic Panel Data Models

- The Keane and Runkle (1992) approach uses the idea of forward filtering from the time-series literature to increase efficiency.
- It involves obtaining a consistent estimate of $\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} \hat{U}^i \hat{U}^{i'}$, where $\hat{U}^i$ is the residual vector for individual i from 2SLS estimation.
- Then it calculates $\hat{Q} = (I_N \otimes \hat{P})$ where $\hat{P}$ is the upper-triangular Cholesky decomposition of $\hat{\Sigma}^{-1}$.
- Finally, it transforms (1) or (3) by premultiplying the equation with $\hat{Q}$ and runs 2SLS on the transformed equation using the original instrument set. This will yield the following slope coefficients:

$$\hat{\beta}_{KR} = (X'\hat{Q}' Z(Z'Z)^{-1} Z'\hat{Q}X)^{-1} X'\hat{Q}' Z(Z'Z)^{-1}\hat{Q}Y. \qquad (4)$$

- Key Point: Pre-multiplication of (3) by $\hat{Q}$ makes the transformed residual for time t a function of epsilons dated $t-1, t, t+1, ..., T$, so the original instruments, such as y_{it-2} and $\boldsymbol{x}_{it-2}$, remain valid.

The XTKR Command

- XTKR implements the Keane and Runkle (1992) approach to panel data models.
- Syntax:

  xtkr *depvar* [*varlist1*] (*varlist2* = *varlist3*) [*if*] [*in*] [, nocons tdum]
- *varlist1* refers to any exogenous variables, *varlist2* refers to the endogenous variables, and *varlist3* refers to the excluded instruments.
- The option ‘tdum’ demeans the data across the time dimension prior to estimation. This is equivalent and preferable to adding time dummies to the regression as that can cause collinearity in the second stage of the estimator.
- It can be installed through ssc by simply typing ‘ssc install xtkr’.

Empirical Example

- We will now replicate an empirical example in Baltagi (2005) that shows the results of these estimators to the dynamic demand for cigarettes in the US from 1963 to 1992.
- The specified equation is:

$$lnC_{it} = \alpha + \beta_1 ln(C_{i,t-1}) + \beta_2 ln(P_{it}) + \beta_3 ln(Pn_{it}) + \beta_4 ln(Y_{it}) + u_{it} \quad (5)$$

- where i is a US state, $N = 46$, and $T = 30$. We assume the three regressors are predetermined and that their lagged values are used to instrument for the lag of consumption.

Empirical Example

Estimator	lnC_{it-1}	lnP_{it}	$lnPn_{it}$	lnY_{it}	Instruments
OLS	0.954	-0.137	0.037	-0.009	N/A
	(148.5)	(-8.67)	(2.72)	(-1.13)	
SYS-GMM	0.942	-0.172	0.047	0.002	466
	(123.85)	(-9.51)	(3.19)	(0.29)	
DIFF-GMM	0.843	-0.377	-0.016	0.139	437
	(52.66)	(-11.81)	(-0.39)	(3.88)	
2SLS-DIFF	0.645	-0.406	0.038	0.156	6
	(4.24)	(-11.77)	(0.86)	(2.84)	
KR-DIFF	0.703	-0.338	0.075	0.225	6
	(17.52)	(-13.51)	(2.59)	(6.44)	

Note: t-statistics are in parenthesis. All regressions include time fixed effects.

Monte Carlo Simulations

- In this part we will test the Keane and Runkle (1992) approach against DIFF-GMM and SYS-GMM using Monte Carlo simulations.
- We will also consider two proposals to avoid the many-instruments problem in the GMM estimators:
  - restricting the number of lags in the instrument set, and
  - collapsing the instrument matrix.
- The data-generating process is:

$$y_{it} = \beta_0 + \beta_1 y_{it-1} + \beta_2 x_{it} + \mu_i + \epsilon_{it} \tag{6}$$

- where the regressor follows the process:

$$x_{it} = \eta_i + 0.5\mu_i + 0.5\epsilon_{it} + 0.5\epsilon_{it-1} + \omega_{it} \tag{7}$$

- μ_i, η_i, ϵ_{it}, and ω_{it} are generated IID.
- $\beta_1 = 0.5$ and $\beta_2 = 1$.

Monte Carlo Simulations

(N=100,T)		**Bias (x100)**			**RMSE (x100)**		
Instr. Set	Instr. Count	5	10	20	5	10	20
Results for $\beta_1 = 0.5$							
FE-OLS	N/A	-19.33	-10.71	-7.39	19.56	10.82	7.45
KR-DIFF							
3 lags	4	-3.09	-0.53	-0.20	15.22	5.15	2.80
4 lags	6	-9.11	-1.13	-0.34	23.56	5.18	2.70
DIFF-GMM							
full	12/72/342	-6.78	-5.85	-5.46	10.65	6.54	5.63
full (collapse)	6/16/36	-6.81	-3.27	-1.86	12.99	5.20	2.73
2 lags	10/30/70	-5.94	-4.03	-3.40	10.46	5.59	4.07
2 lags (collapse)	4	-5.08	-0.71	-0.13	15.83	6.37	3.17
SYS-GMM							
full	19/89/379	0.21	2.53	4.66	6.44	4.23	5.21
full (collapse)	9/19/39	-5.63	-2.82	-1.67	11.10	4.86	2.61
2 lags	17/47/107	0.20	1.25	1.66	6.65	3.78	2.86
2 lags (collapse)	7	-4.40	-0.58	-0.08	11.88	5.49	3.13

Monte Carlo Simulations

(N=100,T)		**Bias (x100)**			**RMSE (x100)**		
Instr. Set	Instr. Count	5	10	20	5	10	20
Results for $\beta_2 = 1$							
FE-OLS	N/A	25.03	32.61	34.60	25.43	32.73	34.65
KR-DIFF							
3 lags	4	1.53	0.99	0.39	23.40	11.92	7.87
4 lags	6	4.75	2.68	0.86	31.66	12.12	7.66
DIFF-GMM							
full	12/72/342	11.95	18.79	24.99	20.72	20.05	25.24
full (collapse)	6/16/36	20.74	15.15	10.93	40.03	21.07	13.39
2 lags	10/30/70	10.79	12.37	12.54	21.36	15.58	13.91
2 lags (collapse)	4	18.08	3.20	0.71	56.80	24.90	11.05
SYS-GMM							
full	19/89/379	11.97	16.01	18.65	20.36	18.12	19.49
full (collapse)	9/19/39	14.74	9.74	7.37	28.95	14.95	9.82
2 lags	17/47/107	10.82	11.74	11.59	20.27	15.11	13.16
2 lags (collapse)	7	11.35	2.36	0.69	31.76	15.39	8.96

Conclusion

- The KR approach has two key features relative to other popular panel-data estimators:
    - It relies on a small instrument set (typically one or two lags of the predetermined variables).
    - It relies on forward filtering to eliminate serial correlation.
- We show in an application in Cigarette demand that the KR estimator generated coefficients that are arguably more theoretically plausible than alternative methods.
- We also report some MC simulation results that show that the KR approach can potentially perform better than the alternatives, consistent with the findings of Ziliak (1997).
- Now that the estimator is available in Stata through XTKR and has simple syntax, we believe it should be in the toolkit of applied researchers.