Testing for Unit Roots in Panel Data:

An Exploration Using Real and Simulated Data

Bronwyn H. HALL

UC Berkeley, Oxford University, and NBER

Jacques MAIRESSE

INSEE-CREST, EHESS, and NBER

Introduction

- Our Research Program:
  - Develop simple models that describe the time series behavior of key variables for a panel of firms:
    - *Sales, employment, profits, investment, R&D*
    - *U.S., France, Japan*
  - **Substantive interest:** use of these variables for further modeling (productivity, investment, etc.) requires an understanding of their univariate behavior
  - **Technical interest:** explore the use of a number of estimators and tests that have been proposed in the literature, using real data.
- This paper: a comparison of unit root tests for fixed T, large N panels, using DGPs that mimic the behavior of our real data.

Outline

- Basic features of our data
- Motivation – issues in estimating a simple dynamic panel model
- Overview of unit root tests for short panels
- Simulation results
- Results for real data

Dataset Characteristics Scientific Sector, 1978-1989

Country	**France**	**United States**	**Japan**
Data sources	Enquete annuelle sur les moyens consacres a la recherche et au dev. dans les entreprises;enq. annuelle des entreprises	Standard and Poor's Compustat data – annual industrial and OTC OTC, based on 10-K filings to SEC	Needs data; Data from JDB (R&D data from Toyo Keizai survey)
# firms	953	863	424
# observations	5,842	6,417	5,088
After cleaning	5,139	5,721	4,260
No jumps	5,108	5,312	4,215
Balanced 1978-89			
(# obs.)	**1,872**	**2,448**	**2,652**
(# firms)	**156**	**204**	**221**
Positive Cash Flow			
(# firms)	**104**	**174**	**200**

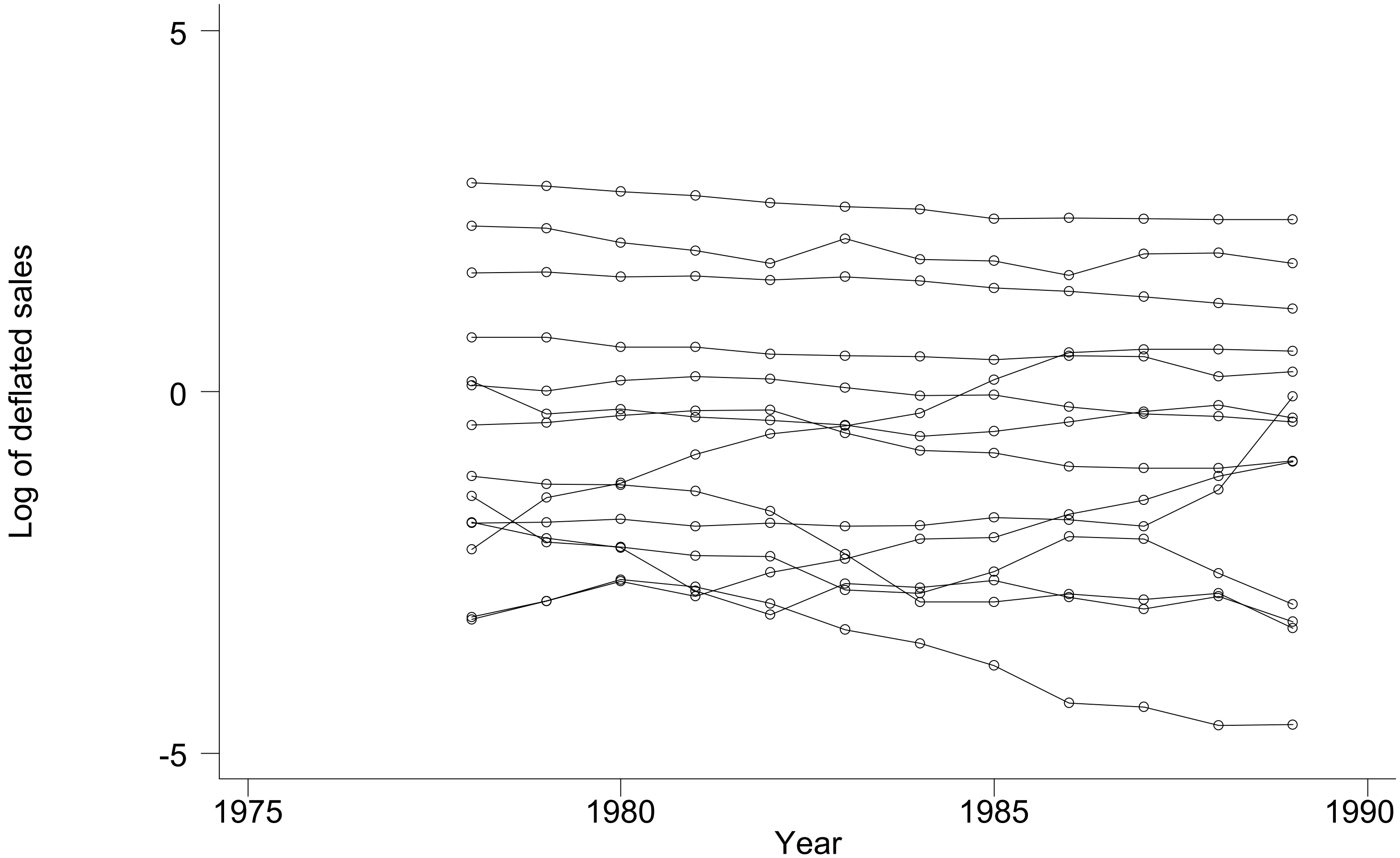
The scientific sector consists of firms in Chemicals, Pharmaceuticals, Electrical Machinery, Computing Equipment, Electronics, and Scientific Instruments.

Variables

- Sales (millions $)
- Employment (1000s)
- Investment (P&E, millions $)
- R&D (millions $)
- Cash flow (millions $)

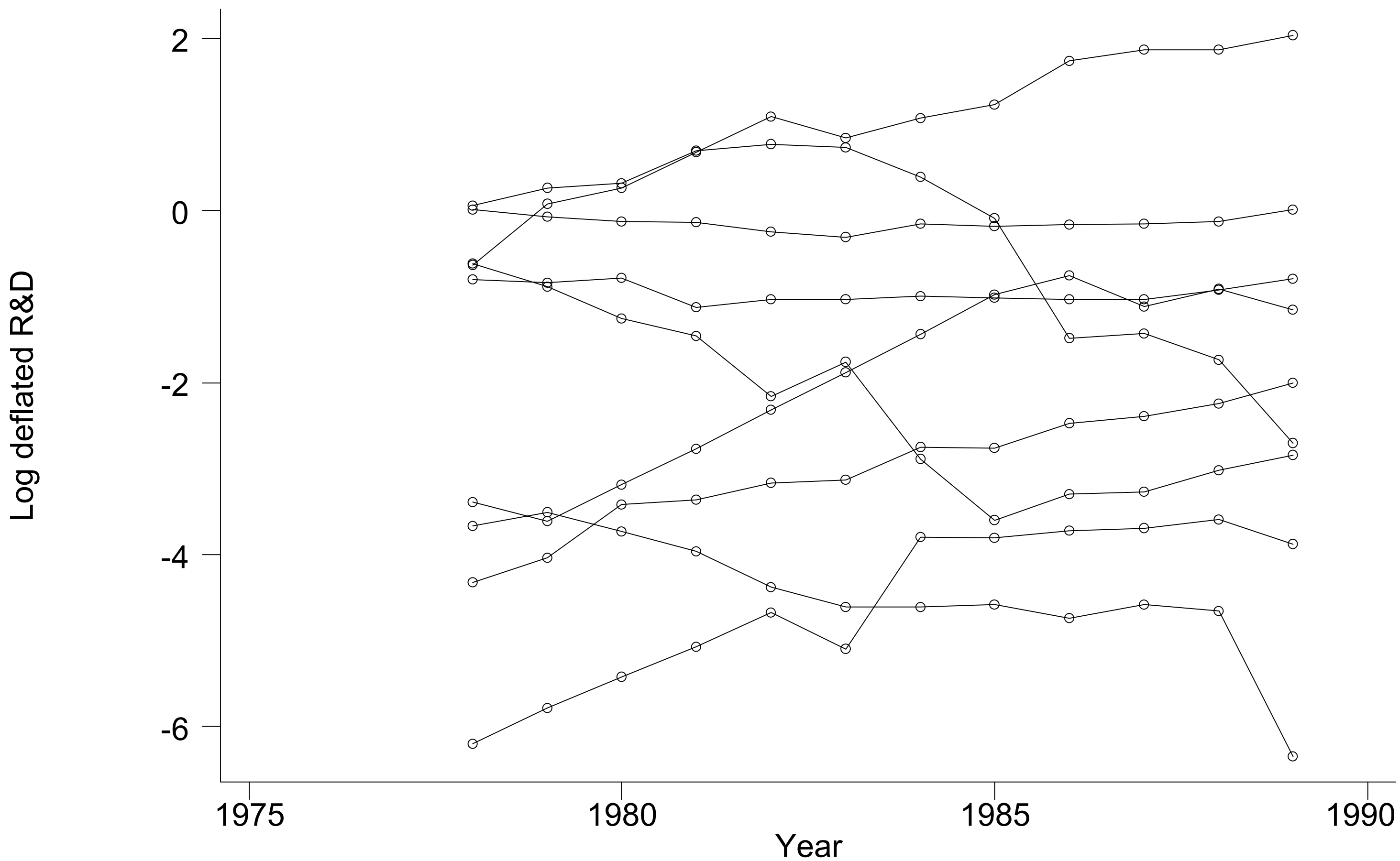
All variables in logarithms, overall year means removed (so price level changes common to all firms are removed – Levin and Lin 1993).

Representative data - sales

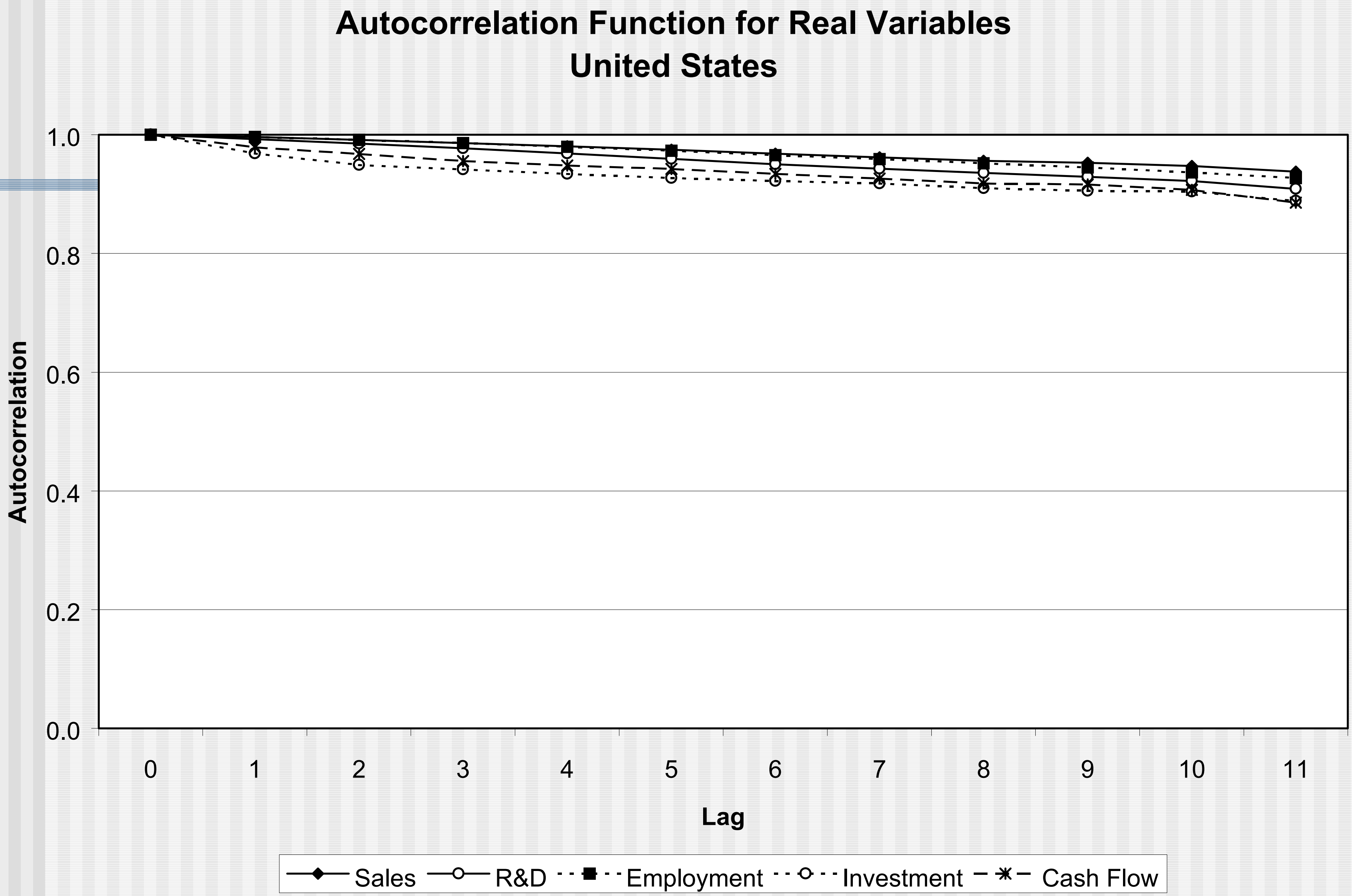


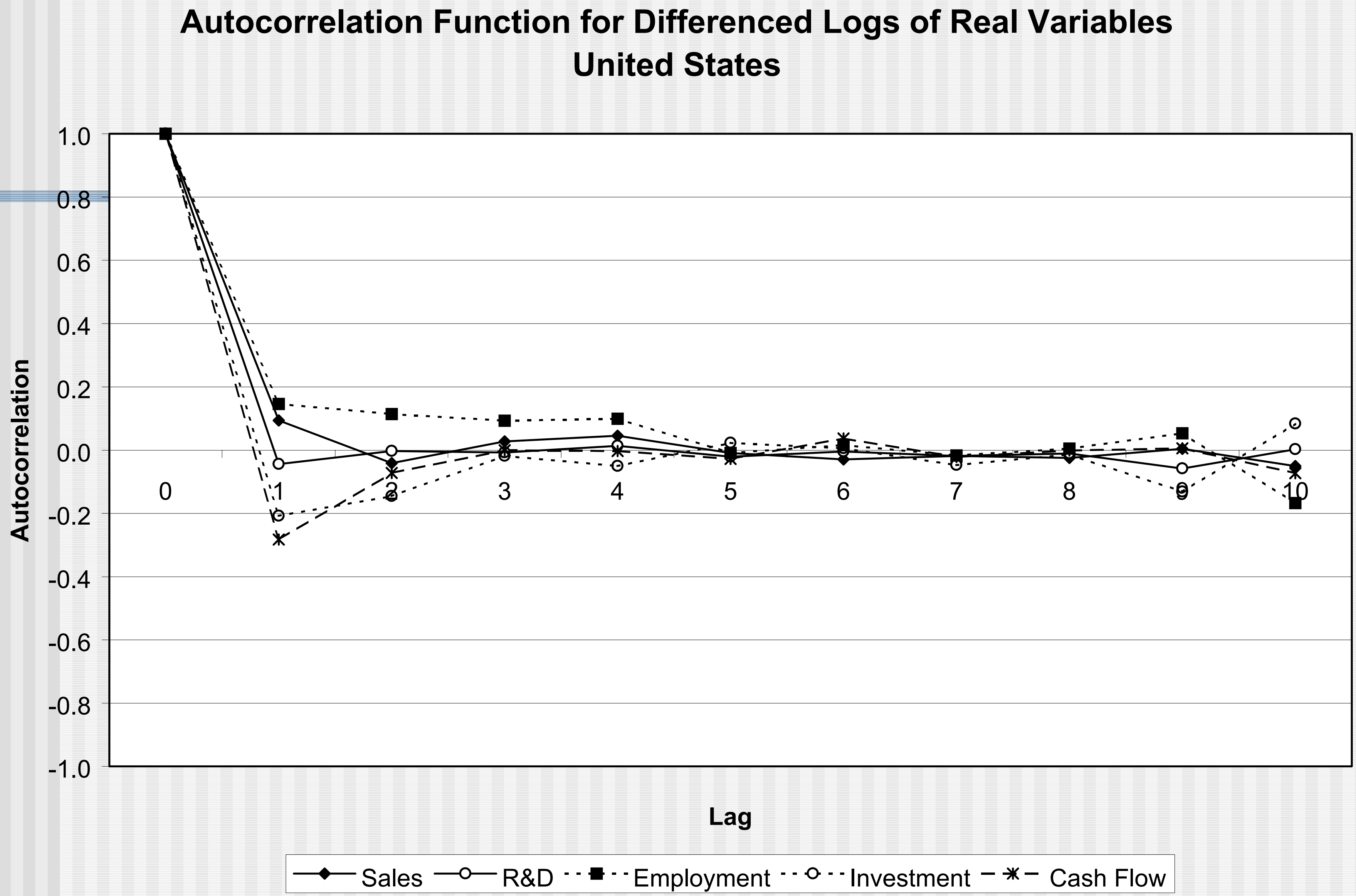
Selected U.S. Manufacturing Firms

Representative data – R&D



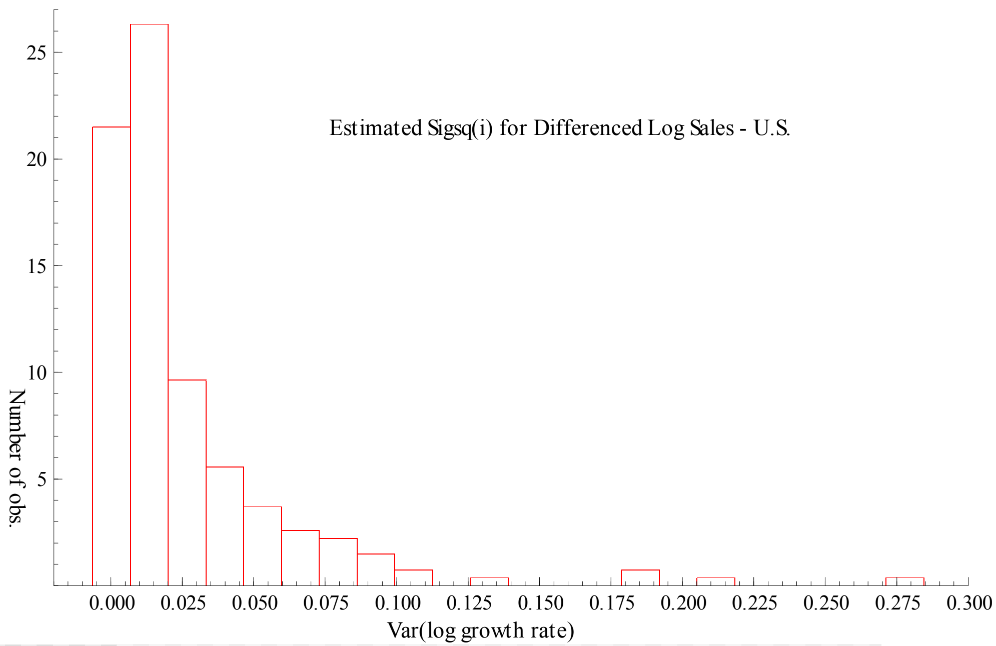
Selected U.S. Manufacturing Firms

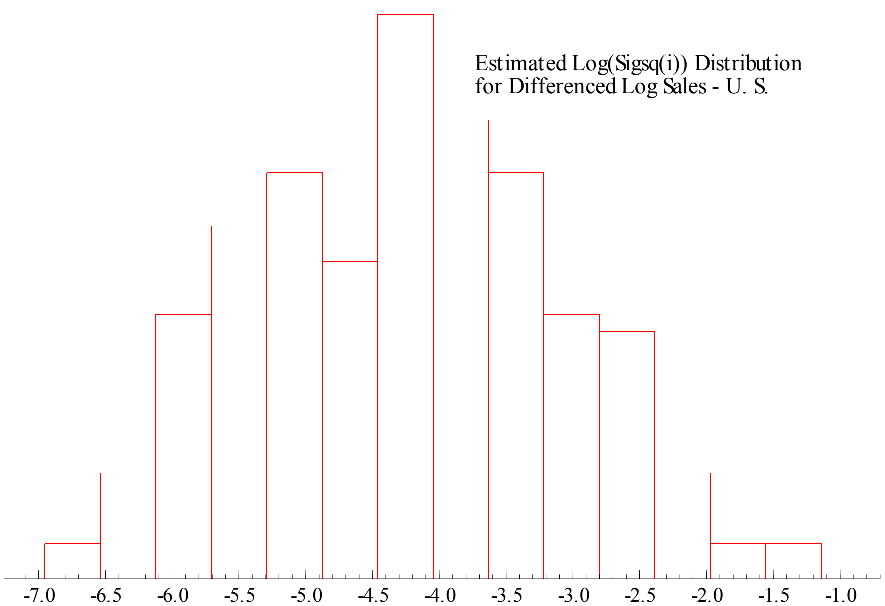

Autocorrelation Function for Real Variables
United States
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8
9
10
11
Lag
Sales
R&D
Employment
Investment
Cash Flow


Autocorrelation Function for Differenced Logs of Real Variables
United States
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
0
1
2
3
4
5
6
7
8
9
10
Lag
Sales
R&D
Employment
Investment
Cash Flow

Variance of Log Growth Rates

$\sigma^2(i)$ | $\log \sigma^2(i)$





Summary

1. Substantial heterogeneity in levels and variances across firms.
   - However, firm-by-firm estimations yield trends with distributions similar to those expected due to sampling error when T is small. (not shown)
   - The sigma-squared distribution differs from that predicted by sampling error, implying heteroskedasticity. (*see graph*)
2. High autocorrelation in levels => fixed effects or autoregression with root near one?
3. Very slight autocorrelation in differences; however, the within coefficient is substantial and positive =>heterogeneity in growth rates?

A Simple Model

y_{it} = logarithm of the variable of interest.

$$y_{it} = \alpha_i + \delta_t + u_{it}$$

$$u_{it} = \rho u_{it-1} + \varepsilon_{it}$$

$i = 1,\ldots,N$ Firms; $t = 1,\ldots,T$ Years

$\varepsilon_{it} \sim (0,\sigma_i^2)$ $\qquad E[\varepsilon_{it}\varepsilon_{js}] = 0, t \neq s$ or $j \neq i$

$$y_{it} = \alpha_i(1-\rho) + \delta_t - \rho\delta_{t-1} + \rho y_{i,t-1} + \varepsilon_{it}$$

$$\Rightarrow (FE): y_{it} = (1-\rho)(\alpha_i + \delta_t) + \rho(\Delta\delta_t + y_{i,t-1}) + \varepsilon_{it}$$

$$\Rightarrow (RW): y_{it} = \Delta\delta_t + y_{i,t-1} + \varepsilon_{it} \quad \text{if} \quad \rho = 1$$

Estimation with a Firm Effect

Drop δ_t (means removed) and difference out α_i:

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta \varepsilon_{it}$$

OLS is inconsistent; use IV or GMM-IV for estimation with $y_{i,t-2},\ldots,y_{i1}$ as instruments.

Advantages: *robust* to heteroskedasticity and non-normality; *consistent* for β's; allows for some types of *transitory* measurement error in *y*.

Disadvantages: *biased* in finite samples; *imprecise* when instruments are weakly correlated with independent variables.

Three Data Generating Processes

1. $\rho \equiv 1 \Rightarrow y_{it} = y_{i,t-1} + \delta + \varepsilon_{it}$

   or $\Delta y_{it} = \delta + \varepsilon_{it}$

OLS is consistent; IV with lagged instruments not identified.

2. $\rho = 0 \Rightarrow y_{it} = \alpha_i + \delta t + \varepsilon_{it}$

   or $\Delta y_{it} = \delta + \Delta\varepsilon_{it}$

OLS is inconsistent; IV or GMM with lag 2+ inst. is consistent

3. $\rho < 1$, no effects $\Rightarrow y_{it} = \alpha + \rho y_{i,t-1} + \delta t + \varepsilon_{it}$

   or $\Delta y_{it} = \rho \Delta y_{i,t-1} + \delta + \Delta\varepsilon_{it}$

OLS is inconsistent; IV or GMM with lag 2+ inst. is consistent

Results of Simulation

N=200 T=12 No. of draws=1000
Estimated coefficient for dy on dy(-1)
Instruments are y(-2)-y(-4)

Truth	OLS	IV	GMM1	GMM2	GMM CUE
rho=1.0 (RW)	-0.001 (.026)	0.279 (.690)	-0.040 (.175)	0.440 (.228)**	-0.047 (.168)
rho=0.0 (FE)	-0.500 (0.019)**	0.000 (.046)	-0.028 (0.042)	-0.010 (.333)	-0.006 (.041)
rho=0.9 (no effects)	-0.059 (.025)**	0.868 (.089)			

** Different from truth at 5% level of significance.

Conclusion from Simulations

- As with ordinary times series, it is essential to test first for a unit root (even though asymptotics in the panel data case are for N and not T).
- Failure to do so may lead to the use of estimators that are very biased and misleading in finite samples even though they are consistent.
  - If unit root => assume no fixed effect and then OLS level estimators appropriate.
  - If no unit root => fixed effect (usually) and IV.
  - Near unit root => OLS bias can be large.

Unit Root Tests Considered

Note that these tests are generally valid for large N and fixed T.

- **IPS**: *Im, Pesaran, and Shim (1995)* – alternative is ρ_i <1 for some *i.* Based on an average of augmented Dickey-Fuller tests conducted firm by firm, with or without trend. Normal disturbances assumed.
- **HT***: Harris-Tzavalis (JE 1999)* – alternative is ρ<1. Based on the LSDV estimator, corrected for bias and normalized by the theoretical std. error under the null. Homoskedastic normal disturbances assumed.

Unit Root Tests (continued)

- **SUR**: *OLS with no fixed effects and an equation for each year* (suggested by Bond et al 2000) – consistent under the null of a unit root. Has good power. Allows for heteroskedasticity and correlation over time easily.
- **CMLE**:
  - *Kruiniger (1998, 1999)* – CMLE is consistent for stationary model and for $\rho=1$ (fixed T). Use an LR test based on this fact. Homoskedastic normal disturbances assumed, but not necessary.
  - *Lancaster and Lindenhovius (1996); Lancaster (1999)* – similar to Kruiniger. Bayesian estimation with flat prior on effects and $1/\sigma$ for the variance yields estimates that are consistent when $\rho=1$ (fixed T). σ is shrunk slightly toward zero.
  - **CMLE-HS**: suggested in *Kruiniger (1998)* – heteroskedasticity of the form $\sigma_i^2\ \sigma_t^2$ can be estimated consistently.

Conditional ML Estimation (HS)

Model: $y_{it} = (1-\rho)\alpha_i + \rho y_{i,t-1} + \varepsilon_{it}$

Or $y_{it} = \alpha_i + u_{it}$

$u_{it} = \rho u_{i,t-1} + \varepsilon_{it} \quad \varepsilon_{it} \sim N(0,\sigma_i^2)$

Stacking the model: $y_i = \alpha_i \iota + u_i$

With $E[u_i u_i'] = \sigma_i^2 V_\rho = \dfrac{\sigma_i^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \dots & \rho^{T-1} \\ \rho & 1 & \dots & \rho^{T-2} \\ \rho^2 & \rho & \dots & \rho^{T-3} \\ \dots & \dots & \dots & \dots \\ \rho^{T-1} & \rho^{T-2} & \dots & 1 \end{bmatrix}$

Conditional ML Estimation (HS)

Differenced:

$$Dy_i = Du_i \quad \text{where} \quad D = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ 0 & 0 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

$$\Rightarrow Dy_i \sim N(0, \Sigma) \text{ with } \Sigma = \sigma_i^2 D V_\rho D' = \sigma_i^2 \Phi$$

The log likelihood function:

$$\log L(\rho, \{\sigma_i^2\}) = \frac{-N(T-1)}{2}\log(2\pi) - \frac{(T-1)}{2}\sum_{i=1}^{N}\log(\sigma_i^2)$$

$$-\frac{N}{2}\log|\Phi| - \frac{1}{2}\sum_{i=1}^{N}\frac{(Dy_i)'\Phi^{-1}Dy_i}{\sigma_i^2}$$

Conditional ML Estimation (HS)

The σ_i^2 can be concentrated out using

$$\sigma_i^2 = \frac{1}{T-1} tr\left(\Phi^{-1} Dy_i (Dy_i)'\right)$$

which yields

$$\log L(\rho) = \frac{-N(T-1)}{2} \log(2\pi + 1) - \frac{(T-1)}{2} \sum_{i=1}^{N} \log(\sigma_i^2(\rho)) - \frac{N}{2} \log\left|\Phi(\rho)\right|$$

for estimation.

Conditional ML Estimation (HS)

- Kruiniger (1999) proves consistency of the CMLE-HS estimator for $\rho \subset (-1,1]$.
- However, the concentrated or profile likelihood version is problematic:
  - Nuisance parameters (σ_i^2) increase with N – standard error estimates biased downward; not efficient (see B-N & Cox, ex. 4.3).
  - Non-orthogonal parameters (ρ, σ_t^2, and σ_i^2)
- Possible alternatives:
  - Modified profile likelihood - Barndorff-Nielsen and Cox (1994), but not clear how to do this.
  - Integrated likelihood (Woutersen 2000).

Results of Simulations

- **IPS**
  - zero augmenting lags to be consistent with other tests.
  - we found size was too large if the data were allowed to choose the number of augmenting lags.
  - size slightly too large
  - power weak against large rho alternatives.
- **HT**
  - size correct if homoskedastic;
  - power weak against large rho alternatives, with or without FE.
- **SUR**
  - size correct; slightly too large if heteroskedastic
  - power weak against large rho alternatives, with or without FE.

Results of Simulations

- **CMLE**
  - size correct if homoskedastic
  - power weak against large rho alternatives, with or without FE
- **CMLE-HS**
  - size wrong
  - power slightly weak against large rho alternatives, with or without FE
  - requires sandwich var-cov estimator; appears to have downward-biased standard errors, so rejects too often.

Results of Simulation - Homoskedastic DGP

N=200   T=12   No. of draws=1000
Empirical size or power (nominal size=.05)

Truth (DGP)	IPS no trend	IPS trend	H-T	CMLE t test	CMLE-HS t test	SUR
rho=1.0 (RW)	.067	.100	.062	.056	.520	.073
rho=0.0 (FE)	1.00	1.00	1.00	1.00	1.00	1.00
rho=0.99 (no effects)	.486	.125	.193	.260	.520	.370

Results of Simulation - Heteroskedastic DGP

N=200   T=12   No. of draws=1000
Empirical size or power (nominal size=.05)

Truth (DGP)	IPS no trend	IPS trend	H-T	CMLE t test	CMLE-HS t test	SUR
rho=1.0 (RW)	.090	.050	.210	.200	.450	.124
rho=0.0 (FE)	1.00	1.00	1.00	1.00	1.00	1.00
rho=0.99 (no effects)	.125	.240	.369	.390	.550	.303

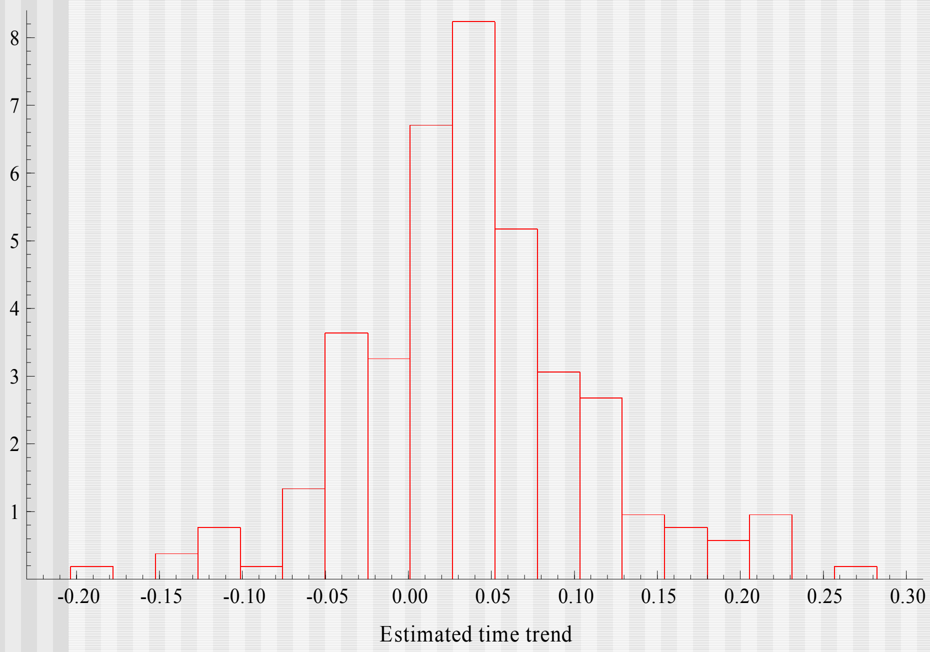
Results of Unit Root Tests
Series with unit roots

	IPS no trend	IPS with trend	HT	CMLE	CMLE with HS	SUR
Sales	US,J	US,F,J	US,F,J	US,F,J	US,F,J	J only
Employment	US,F,J	US,F,J	US,F,J	US,F	US,F,J	J only
R&D	US only	US,F,J	US only	US only	US,F,J	--
Investment	--	--	--	--	--	--
Cash flow	US only	US,J	--	--	--	--

Conclusions

- A model with a very large autoregressive coefficient and no level fixed effect may be a good description of these data – the substantive implication is that we use the initial condition rather than a permanent "effect" to describe differences across firms.
- CML estimation is feasible and may be a useful estimator in the cases where we cannot use the SUR idea.
- Next steps:
  - Heteroskedastic-consistent standard errors to correct size in CMLE-HS, etc.
  - Further exploration of heterogeneous trends.
  - Modeling a more complex AR process for our data with heteroskedasticity but no fixed effects.

Trends – real and simulated data



Intercepts – real and simulated data