

# Longitudinal/Panel Data Analysis: Lecture 3 and 4

Raymond Duch

University of Oxford  
Nuffield College  
[raymond.duch@nuffield.ox.ac.uk](mailto:raymond.duch@nuffield.ox.ac.uk)  
[raymond Duch.com/class/paneldata](http://raymond Duch.com/class/paneldata)

May 26, 2008

- Stata 10.0 Manual Longitudinal/Panel Data, xtabond, xtabond postestimation, xtdpdsys, xtivreg
- Green, Donald P. and David H. Yoon. 2002. "Reconciling Individual and Aggregate Evidence Concerning Partisan Stability: Applying Time-Series Models to Panel Survey Data." Political Analysis 10:1-24.
- Wawro, Gregory. 2002. "Estimating Dynamic Panel Data Models in Political Science." Political Analysis 10:25-48.
- Halaby, Charles N. 2004. "Panel Models in Sociological Research: Theory into Practice." Annual Review of Sociology. 30:507-44.
- Beck, Nathaniel and Jonathan N. Katz. 2007. "Random Coefficient Models for Time-Series-Cross-Section Data: Monte Carlo Experiments." Political Analysis. 15: 182-195.

- Wilson, Sven E. and Daniel M. Butler. 2007. "A Lot More to Do: The Sensitivity of Time-Series Cross-Section Analyses to Simple Alternative Specifications." *Political Analysis*. 15:101-123.
- Plumper, Thomas and Vera E. Troeger. 2007. "Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects." *Political Analysis* 15:124-139.
- Shor, Boris, Joseph Bafumi, Luke Keele and David Park. 2007. "A Bayesian Multilevel Modeling Approach to Time-Series Cross-Sectional Data." *Political Analysis*. 15:165-181.

$$y_{jt} = \alpha + \sum_k \beta_k \omega_{kjt} + \sum_p \phi_p z_{jp} + \gamma x_{jt} + \theta_j + \epsilon_{jt} \quad (1)$$

- $j = 1, \dots, N$   $t=1, \dots, T$
- $\theta_j$  is a term for unit effects
- causal variable of interest is  $x_{jt}$
- $\omega_{kjt}$  ( $k = 1, \dots, K$ ) consists of additional explanatory variables that vary over time
- $z_{jp}$  consists of additional explanatory variables that do not vary over time

- Should  $\theta_j$  be treated as random or fixed?
- if unobserved  $\theta_j$  are uncorrelated with regressors random effects is reasonable
- serial correlation of composite errors suggests efficiency gains from GLS
- but if unit effects correlated with explanatory variables estimators may be biased and inconsistent

$$(y_{jt} - y_{jt-1}) = \sum_k \beta_k (\omega_{kjt} - \omega_{kjt-1}) + \gamma (x_{jt} - x_{jt-1}) + (\epsilon_{jt} - \epsilon_{jt-1}) \quad (2)$$

$$(y_{jt} - \bar{y}_j) = \sum_k \beta_k (\omega_{kjt} - \bar{\omega}_{kj}) + \gamma (x_{jt} - \bar{x}_j) + (\epsilon_{jt} - \bar{\epsilon}_j) \quad (3)$$

If the disturbances in the transformed equations are constant variance and serially uncorrelated then both fixed effects and difference estimators are efficient for a fixed effects model.

$$y_{j1} = \delta_1 + \phi_1 z_j + \gamma(x_{j1}) + \theta_j + \epsilon_{j1} \quad (4)$$

$$y_{j2} = \delta_2 + \phi_2 z_j + \gamma(x_{j2}) + \theta_j + \epsilon_{j2} \quad (5)$$

$$(y_{j2} - y_{j1}) = (\delta_2 - \delta_1) + (\phi_2 - \phi_1)z_j + \gamma(x_{j2} - x_{j1}) + (\epsilon_{j2} - \epsilon_{j1}) \quad (6)$$

The change in  $(\phi_2 - \phi_1)$  is estimable as are time-invariant variables interacted with time trends or periods.

$$y_{jt} = \alpha + \sum_k \beta_k \omega_{kjt} + \sum_p \phi_p z_{jp} + \gamma x_{jt} + \alpha_j + \epsilon_{jt} \quad (7)$$

We can estimate  $\gamma$  employing either a fixed effect ( $\gamma_{fe}$ ) estimator or a random effect ( $\gamma_{re}$ ) estimator. Hausman showed that  $(\hat{\gamma}_{fe} - \hat{\gamma}_{re})$  could be used to test the null hypothesis that the unit effects and the explanatory variables are uncorrelated.

Small values of Hausman Test indicate a failure to reject null hypothesis that the unit effects and the explanatory variables are uncorrelated and favour the random effects estimation of fixed effects models.

$$y_{jt} = \phi_1 z_{1j} + \phi_2 z_{2j} + \gamma_1 x_{1jt} + \gamma_2 x_{2jt} + \theta_j + \epsilon_{jt} \quad (8)$$

- all explanatory variables are strictly exogenous
- $z_{2j}$  and  $x_{2jt}$  (but not  $z_{1j}$  and  $x_{1jt}$ ) are correlated with  $\theta_j$
- unbiased estimates of  $\phi_1$  and  $\gamma_1$  is unproblematic because  $z_{1j}$  and  $x_{1jt}$  are exogenous and therefore act as their own instruments
- $(x_{2jt} - \bar{x}_{2j})$  is an instrument for  $x_{2jt}$
- $\bar{x}_{1j}$  is an instrument for  $z_2$

. xthtaylor lwage occ south smsa ind exp exp2 wks ms union fem blk ed, endog(exp exp2 wks ms union ed)

```

Hausman-Taylor estimation
Group variable: id
Number of obs      =      4165
Number of groups   =      595
Obs per group: min =         7
                  avg  =         7
                  max  =         7
Random effects u_i ~ i.i.d.
Wald chi2(12)      =     6891.87
Prob > chi2        =      0.0000

```

lwage	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
TVexogenous						
occ	-.0207047	.0137809	-1.50	0.133	-.0477149	.0063055
south	.0074398	.031955	0.23	0.816	-.0551908	.0700705
smsa	-.0418334	.0189581	-2.21	0.027	-.0789906	-.0046761
ind	.0136039	.0152374	0.89	0.372	-.0162608	.0434686
TVendogenous						
exp	.1131328	.002471	45.79	0.000	.1082898	.1179758
exp2	-.0004189	.0000546	-7.67	0.000	-.0005259	-.0003119
wks	.0008374	.0005997	1.40	0.163	-.0003381	.0020129
ms	-.0298508	.01898	-1.57	0.116	-.0670508	.0073493
union	.0327714	.0149084	2.20	0.028	.0035514	.0619914
Tlexogenous						
fem	-.1309236	.126659	-1.03	0.301	-.3791707	.1173234
blk	-.2857479	.1557019	-1.84	0.066	-.5909179	.0194221
Tlendogenous						
ed	.137944	.0212485	6.49	0.000	.0962977	.1795902
_cons	2.912726	.2836522	10.27	0.000	2.356778	3.468674
-----+-----						
sigma_u	.94180304					
sigma_e	.15180273					
rho	.97467788	(fraction of variance due to u_i)				
-----+-----						

Note: TV refers to time varying; TI refers to time invariant.

# Features of Dynamic Panel Models

- ① include in their specification 1) lagged dependent variables and 2) unobserved individual-specific effects
- ② models are powerful because allow for 1) empirical modeling of dynamics while 2) accounting for individual-specific dynamics
- ③ parse out: 1) does past behaviour directly affect current behaviour (dynamic effect) versus 2) individuals have predilection to behave in particular way (individual-specific)

$$y_{jt} = \varphi y_{jt-1} + \phi z_j + \gamma x_{jt} + \theta_j + \epsilon_{jt} \quad (9)$$

- all explanatory variables are strictly exogenous
- $\epsilon_{jt}$  is mean zero, constant variance, and independently distributed
- least squares estimator is badly biased because of the correlation between the lagged endogenous variable  $y_{jt-1}$  and the unit effects – because  $\epsilon_{jt}$  affects  $y_{jt}$  it also affects  $y_{jt-1}$

$$(y_{jt} - y_{jt-1}) = \varphi(y_{jt-1} - y_{jt-2}) + \sum_k \beta_k (x_{kjt} - x_{kjt-1}) + (\epsilon_{jt} - \epsilon_{jt-1}) \quad (10)$$

- eliminate unit effect by first differencing the variables
- apply instrumental variables estimation for the parameters of the lagged endogenous variable ( $\varphi$ )
- Anderson and Hsaio use  $y_{jt-2}$  as an instrument for  $(y_{jt-1} - y_{jt-2})$

# Stata Instrumental Variable Estimation (xtivreg)

```
. xtivreg n l2.n l(0/1).w l(0/2).(k ys) yr1981-yr1984 (l.n=l3.n), fd
```

First-differenced IV regression

Group variable:	id	Number of obs	=	471
Time variable (t):	year	Number of groups	=	140
R-sq: within	= 0.0141	Obs per group: min	=	3
between	= 0.9165	avg	=	3.4
overall	= 0.9892	max	=	5
		Wald chi2(14)	=	122.53
corr(u_i, Xb)	= 0.9239	Prob > chi2	=	0.0000

# Stata Instrumental Variable Estimation(xtivreg)

D.n	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
n						
LD.	1.422765	1.583053	0.90	0.369	-1.679962	4.525493
L2D.	-.1645517	.1647179	-1.00	0.318	-.4873928	.1582894
w						
D1.	-.7524675	.1765733	-4.26	0.000	-1.098545	-.4063902
LD.	.9627611	1.086506	0.89	0.376	-1.166752	3.092275
k						
D1.	.3221686	.1466086	2.20	0.028	.0348211	.6095161
LD.	-.3248778	.5800599	-0.56	0.575	-1.461774	.8120187
L2D.	-.0953947	.1960883	-0.49	0.627	-.4797207	.2889314
ys						
D1.	.7660906	.369694	2.07	0.038	.0415037	1.490678
LD.	-1.361881	1.156835	-1.18	0.239	-3.629237	.9054744
L2D.	.3212993	.5440403	0.59	0.555	-.745	1.387599
yr1981						
D1.	-.0574197	.0430158	-1.33	0.182	-.1417291	.0268896
yr1982						
D1.	-.0882952	.0706214	-1.25	0.211	-.2267106	.0501203
yr1983						
D1.	-.1063153	.10861	-0.98	0.328	-.319187	.1065563
yr1984						
D1.	-.1172108	.15196	-0.77	0.441	-.4150468	.1806253
_cons	.0161204	.0336264	0.48	0.632	-.0497861	.082027
-----+-----						
sigma_u	.29069213					
sigma_e	.18855982					
rho	.70384993	(fraction of variance due to u_i)				

Instrumented: L.n  
 Instruments: L2.n w L.w k L.k L2.k ys L.ys L2.ys yr1981 yr1982 yr1983 yr1984 L3.n

$$(y_{jt} - y_{jt-1}) = \varphi(y_{jt-1} - y_{jt-2}) + \sum_k \beta_k(x_{kjt} - x_{kjt-1}) + \sum_s \phi_s(w_{sjt} - w_{sjt-1}) + (\epsilon_{jt} - \epsilon_{jt-1}) \quad (11)$$

- $x_{kjt}$  are strictly exogenous covariates
- $w_{sjt}$  are predetermined and endogenous covariates
- eliminate unit effect by first differencing the variables
- apply instrumental variables estimation for the parameters of the lagged endogenous variable ( $\varphi$ )

# The Instrumental Variables in Arellano-Bond GMM Estimation

General form of GMM estimator of  $\hat{\beta}$

$$\hat{\beta} = (\mathbf{X}'\mathbf{Z}\hat{\mathbf{W}}\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z}\hat{\mathbf{W}}\mathbf{Z}'\mathbf{Y}) \quad (12)$$

- $\mathbf{Z}$  consists of the valid instruments in the differenced equation
- the GMM-type instruments are lagged values of  $w_{sjt}$  – predetermined and endogenous covariates
- the standard instruments are the first differences of  $x_{kjt}$  – the exogenous variables

$$n_{it} = \alpha_1 y_{it-1} + \alpha_2 y_{it-2} + \gamma_1 w_{it} + \gamma_2 k_{it} + \gamma_3 y_{sit} + \eta_i + \lambda_t + \epsilon_{it} \quad (13)$$

- $n_{it}$  is log employment in company  $i$  at end of year  $t$
- $k$  is log of gross capital;  $w$  is log of real product wage;  $ys$  is log of industry output
- $\lambda_t$  captures a time effect common to all firms
- sample consists of 140 firms over period 176-1984 – note three cross sections lost in constructing lags

$$\mathbf{z}_i = \begin{bmatrix} n_{i,1} & n_{i,2} & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & \vdots & \Delta x_{i,4} \\ 0 & 0 & n_{i,1} & n_{i,2} & n_{i,3} & \cdots & 0 & \cdots & 0 & \vdots & \Delta x_{i,5} \\ \vdots & & & & & \ddots & & & & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & n_{i,1} & \cdots & n_{i,7} & \vdots & \Delta x_{i,9} \end{bmatrix}$$

# Stata Instrumental Variable Estimation (xtabond)

```
. xtabond n 1(0/1).w 1(0/2).(k ys) yr1980-yr1984 year, lags(2) vce(robust) noconstant
```

```
Arellano-Bond dynamic panel-data estimation   Number of obs       =       611
Group variable: id                             Number of groups    =       140
Time variable: year

Obs per group:   min =         4
                  avg =    4.364286
                  max =         6

Number of instruments =      41           Wald chi2(16)       =    1727.45
                                                Prob > chi2         =      0.0000
```

# Stata Instrumental Variable Estimation(xtabond)

## One-step results

	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
n						
L1.	.6862261	.1445943	4.75	0.000	.4028266	.9696257
L2.	-.0853582	.0560155	-1.52	0.128	-.1951467	.0244302
w						
--.	-.6078208	.1782055	-3.41	0.001	-.9570972	-.2585445
L1.	.3926237	.1679931	2.34	0.019	.0633632	.7218842
k						
--.	.3568456	.0590203	6.05	0.000	.241168	.4725233
L1.	-.0580012	.0731797	-0.79	0.428	-.2014308	.0854284
L2.	-.0199475	.0327126	-0.61	0.542	-.0840631	.0441681
ys						
--.	.6085073	.1725313	3.53	0.000	.2703522	.9466624
L1.	-.7111651	.2317163	-3.07	0.002	-1.165321	-.2570095
L2.	.1057969	.1412021	0.75	0.454	-.1709542	.382548
yr1980	.0029062	.0158028	0.18	0.854	-.0280667	.0338791
yr1981	-.0404378	.0280582	-1.44	0.150	-.0954307	.0145552
yr1982	-.0652767	.0365451	-1.79	0.074	-.1369038	.0063503
yr1983	-.0690928	.047413	-1.46	0.145	-.1620205	.0238348
yr1984	-.0650302	.0576305	-1.13	0.259	-.1779839	.0479235
year	.0095545	.0102896	0.93	0.353	-.0106127	.0297217

## Instruments for differenced equation

GMM-type: L(2/.)n

Standard: D.w LD.w D.k LD.k L2D.k D.ys LD.ys L2D.ys D.yr1980 D.yr1981 D.yr1982 D.yr1983 D.yr1984

Test of overidentifying restrictions is the Sargan test:

$$s = \hat{\epsilon}' \mathbf{Z} \left( \frac{1}{N} \sum_{i=1}^N \mathbf{z}'_i \hat{\epsilon}_i \hat{\epsilon}'_i \mathbf{z}_i \right)^{-1} \mathbf{z}'_i \hat{\epsilon} \quad (14)$$

# Implementing the Sargan test

- 1 Estimating the equation by IV and obtain the residuals  $\hat{\epsilon}_{jt}$
- 2 Regress the IV  $\hat{\epsilon}_{jt}$  on all exogenous variables (instrument + controls)
- 3 Obtain the  $R^2$
- 4 The test statistic is  $S = nR^2$
- 5 Where  $n$  is the number of observations
- 6 Under the null hypothesis that all instruments are exogenous  $S$  is distributed as  $\chi^2_{m-r}$ , where  $m - r$  is the number of instruments minus the number of endogenous variables

# Stata Instrumental Variable Estimation(xtabond)

```
. xtabond n l(0/1).w l(0/2).(k ys) yr1980-yr1984 year, lags(2) noconstant
Arellano-Bond dynamic panel-data estimation Number of obs      =      611
Group variable: id                Number of groups       =      140
Time variable: year

                                Obs per group:   min =         4
                                                avg =      4.364286
                                                max =         6
Number of instruments =      41                Wald chi2(16)         =      1757.07
                                                Prob > chi2           =      0.0000
```

## One-step results

	n	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----							
n							
L1.		.6862261	.1486163	4.62	0.000	.3949435	.9775088
L2.		-.0853582	.0444365	-1.92	0.055	-.1724523	.0017358
w							
--.		-.6078208	.0657694	-9.24	0.000	-.7367265	-.4789151
L1.		.3926237	.1092374	3.59	0.000	.1785222	.6067251
k							
--.		.3568456	.0370314	9.64	0.000	.2842653	.4294259
L1.		-.0580012	.0583051	-0.99	0.320	-.172277	.0562747
L2.		-.0199475	.0416274	-0.48	0.632	-.1015357	.0616408
ys							
--.		.6085073	.1345412	4.52	0.000	.3448115	.8722031
L1.		-.7111651	.1844599	-3.86	0.000	-1.0727	-.3496304
L2.		.1057969	.1428568	0.74	0.459	-.1741974	.3857912
yr1980		.0029062	.0212705	0.14	0.891	-.0387832	.0445957
yr1981		-.0404378	.0354707	-1.14	0.254	-.1099591	.0290836
yr1982		-.0652767	.048209	-1.35	0.176	-.1597646	.0292111
yr1983		-.0690928	.0627354	-1.10	0.271	-.1920521	.0538664
yr1984		-.0650302	.0781322	-0.83	0.405	-.2181665	.0881061
year		.0095545	.0142073	0.67	0.501	-.0182912	.0374002

# Stata Instrumental Variable Estimation: A-B Test for Serial Correlation and Sargan Test

```
. estat abond
```

```
Arellano-Bond test for zero autocorrelation in first-differenced errors
```

```
+-----+  
|Order | z      Prob > z|  
|-----+-----|  
|  1  |-3.9394  0.0001 |  
|  2  |-.54239  0.5876 |  
+-----+  
H0: no autocorrelation
```

```
. estat sargan
```

```
Sargan test of overidentifying restrictions
```

```
H0: overidentifying restrictions are valid
```

```
chi2(25)      = 65.81806
```

```
Prob > chi2   = 0.0000
```

```
.  
end of do-file
```

# Stata Instrumental Variable Estimation not assuming strict exogeneity (xtabond)

Recall

$$E[x_{jt}\epsilon_{js}] = 0 \quad (15)$$

for all  $t, s$

Predetermined variables

$$E[x_{jt}\epsilon_{js}] \neq 0 \quad (16)$$

for  $s < t$

# Stata Instrumental Variable Estimation not assuming strict exogeneity (xtabond)

```
. xtabond n l(0/1).ys yr1980-yr1984 year, lags(2) twostep pre(w,lag(1,.)) ///  
> pre(k, lag(2,.)) vce(robust) noconstant  
Arellano-Bond dynamic panel-data estimation Number of obs      =      611  
Group variable: id          Number of groups       =      140  
Time variable: year  
  
Obs per group:   min =      4  
                  avg =  4.364286  
                  max =      6  
  
Number of instruments =      83          Wald chi2(15)      =    958.30  
                  Prob > chi2          =    0.0000
```

# Stata Instrumental Variable Estimation not assuming strict exogeneity (xtabond)

## Two-step results

	n	Coef.	WC-Robust Std. Err.	z	P> z	[95% Conf. Interval]
n						
L1.		.8580958	.1265515	6.78	0.000	.6100594 1.106132
L2.		-.081207	.0760703	-1.07	0.286	-.2303022 .0678881
w						
--.		-.6910855	.1387684	-4.98	0.000	-.9630666 -.4191044
L1.		.5961712	.1497338	3.98	0.000	.3026982 .8896441
k						
--.		.4140654	.1382788	2.99	0.003	.1430439 .6850868
L1.		-.1537048	.1220244	-1.26	0.208	-.3928681 .0854586
L2.		-.1025833	.0710886	-1.44	0.149	-.2419143 .0367477
ys						
--.		.6936392	.1728623	4.01	0.000	.3548354 1.032443
L1.		-.8773678	.2183085	-4.02	0.000	-1.305245 -.449491
yr1980		-.0072451	.0171163	-0.42	0.673	-.0408839 .0263938
yr1981		-.0609608	.030207	-2.02	0.044	-.1201655 -.0017561
yr1982		-.1130369	.0454826	-2.49	0.013	-.2021812 -.0238926
yr1983		-.1335249	.0600213	-2.22	0.026	-.2511645 -.0158853
yr1984		-.1623177	.0725434	-2.24	0.025	-.3045001 -.0201352
year		.0264501	.0119329	2.22	0.027	.003062 .0498381

## Instruments for differenced equation

GMM-type: L(2/.) .n L(1/.) .L.w L(1/.) .L2.k

Standard: D.y L.D.y D.yr1980 D.yr1981 D.yr1982 D.yr1983 D.yr1984 D.yr