The Stata Journal

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press, and Stata is a registered trademark of StataCorp LP.

The Stata Journal (2005)
5, Number 2, pp. 202–207

Estimation and testing of fixed-effect panel-data systems

J. Lloyd Blackwell, III
Department of Economics
University of North Dakota

Abstract. This paper describes how to specify, estimate, and test multiple-equation, fixed-effect, panel-data equations in Stata. By specifying the system of equations as seemingly unrelated regressions, Stata panel-data procedures worked seamlessly for estimation and testing of individual variable coefficients, but additional routines using `test` were needed for testing of individual equations and differences between equations.

Keywords: st0084, panel data, fixed effect, multiple equations, seemingly unrelated regressions, heteroskedasticity, autocorrelation, contemporaneous correlation, tests of linear combinations

1 Introduction

The motivation to develop a multiple-equation panel-data procedure came from a need to determine whether certain laboratory experiments with multiple panels of laboratory animals produced statistically different results. This suggested that each experiment could be treated as a separate equation in a system of seemingly unrelated regressions (SUR), although it was necessary to apply special restrictions prior to estimation. The SUR model described below is a slightly simplified version of one that was developed to estimate and test pairs of laboratory experiments.

2 Framework

The framework for this paper derives from seemingly unrelated regressions with error components (Baltagi 2001, 105–106). It is then assumed that coefficients of constant terms and quantitative independent variables require restriction across the panels in their equations, while fixed-effect dummies vary by panel (Judge et al. 1988, 456–459). Error structures of both equations below were assumed to be characterized by panel heteroskedasticity, panel autocorrelation, and contemporaneous correlation (HPAC). For simplicity, fixed effects were estimated with the dummy variable regression technique (Wooldridge 2002, 272–274). A simple model for exploration of this situation is

$$y_{1it} = b_{10}x_{1it} + b_{11} + \upsilon_{1i} + \varepsilon_{1it}; i = 1, 2, \ldots, N_1; t = 1, 2, \ldots, T \tag{1}$$

$$y_{2it} = b_{20}x_{2it} + b_{21} + \upsilon_{2i} + \varepsilon_{2it}; i = N_1 + 1, \ldots, N_2; t = 1, 2, \ldots, T \tag{2}$$

Equations here are indicated by first subscripts, while other subscripts indicate panels and time periods, as shown. Thus (1) has N_1 panels, T time periods, and N_1 fixed-effect coefficients, $\upsilon_{1,1}$, $\upsilon_{1,2}$, ..., υ_{1,N_1}. Likewise, (2) has $N_2 - N_1$ panels, T time periods, and the fixed-effect coefficients $\upsilon_{2,1}$, $\upsilon_{2,2}$, ..., υ_{2,N_2}. Thus when (1) and (2) are pooled, we have a total of N_2 panels, with no repetition of time codes within any given panel. This is important because estimation fails when there is multiplicity of time codes within any panel.

The SUR matrix specification for (1) and (2) is

$$\mathbf{y} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \boldsymbol{\varepsilon}$$

where $\mathbf{y} = \left[y_{1,1,1}, \ldots, y_{1,N_1,T}, y_{2,N_1+1,1}, \ldots, y_{2,N_2,T}\right]'$

$\mathbf{X}_1 = \left[c_1, x_1, d_{1,2}, \ldots, d_{1,N_1}\right] \quad \mathbf{X}_2 = \left[c_2, x_2, d_{2,N_1+2}, \ldots, d_{2,N_2}\right]$

$\boldsymbol{\beta}_1 = \left[b^*_{1,0}, b_{1,1}, \upsilon^*_{1,2}, \ldots, \upsilon^*_{1,N_1}\right]' \quad \boldsymbol{\beta}_2 = \left[b^*_{2,0}, b_{2,1}, \upsilon^*_{2,N_1+2}, \ldots, \upsilon^*_{2,N_2}\right]'$

$d_{1,2}, \ldots, d_{1,N_1}$ and $d_{2,N_1+2}, \ldots, d_{2,N_2}$ are fixed-effect dummies

$\boldsymbol{\varepsilon} = \left[\varepsilon_{1,1,1}, \ldots, \varepsilon_{1,N_1,T}, \varepsilon_{2,N_1+1,1}, \ldots, \varepsilon_{2,N_2,T}\right]'$

asterisks indicate composite variables

3 Data setup for estimation

In the experimental data described in the introduction, (1) and (2) each had seven panels. Thus the panels of the system were consecutively numbered 1–14, with panels 1–7 assigned to (1) and panels 8–14 assigned to (2). Deletion of one dummy from each equation then left

$$\mathbf{X}_1 = \left[c_1, x_1, d_2, \ldots, d_7\right] \quad \mathbf{X}_2 = \left[c_2, x_2, d_9, \ldots, d_{14}\right]$$

Extracting x_1 from $\mathbf{X}_1$ and x_2 from $\mathbf{X}_2$, the vector $\mathbf{x}$ was formed by stacking x_1 and x_2 to correspond with the elements of $\mathbf{y}$. For purposes of data entry, however, it seemed advisable to unbold $\mathbf{y}$ and $\mathbf{x}$, since Stata normally uses unbolded lowercase letters in variable names. Thus, the variables `y` and `x`, along with identifiers for panels, time codes, and equations were input into a Stata dataset. Remaining variables were then created and arranged by running the do-file:

```
. tab eq, gen(c)
. gen x1 = c1*x
. gen x2 = c2*x
. tab pnl, gen(d)
. drop d1 d8
```

```
. move x1 c2
. move d2 c2
...
. move d7 c2
```

Table 1 shows a subset of the full dataset described above, created by selecting the first three panels and time periods from each equation, thus placing panels 1–3 in (1) and panels 8–10 in (2). Table 1 merits close scrutiny because the panel identifiers, time codes, equation codes, and contents of $\mathbf{X}_1$ and $\mathbf{X}_2$ accurately reflect the organization and features of the full dataset. For example, examination of $\mathbf{X}_1$ and $\mathbf{X}_2$ reveals that a fixed-effect dummy must be deleted from each to avert block-specific dummy variable traps. Thus estimation requires the `nointercept` option, as in the classic SUR specification (Judge et al. 1988, 470).

Table 1: sample dataset

y	x	pnl	week	eq	c1	x1	d2	d3	c2	x2	d9	d10
116	0	1	0	1	1	0	0	0	0	0	0	0
166	0.693	1	1	1	1	0.693	0	0	0	0	0	0
201	1.099	1	2	1	1	1.099	0	0	0	0	0	0
149	0	2	0	1	1	0	1	0	0	0	0	0
197	0.693	2	1	1	1	0.693	1	0	0	0	0	0
230	1.099	2	2	1	1	1.099	1	0	0	0	0	0
126	0	3	0	1	1	0	0	1	0	0	0	0
151	0.693	3	1	1	1	0.693	0	1	0	0	0	0
209	1.099	3	2	1	1	1.099	0	1	0	0	0	0
114	0	8	0	2	0	0	0	0	1	0	0	0
164	0.693	8	1	2	0	0	0	0	1	0.693	0	0
225	1.099	8	2	2	0	0	0	0	1	1.099	0	0
142	0	9	0	2	0	0	0	0	1	0	1	0
206	0.693	9	1	2	0	0	0	0	1	0.693	1	0
262	1.099	9	2	2	0	0	0	0	1	1.099	1	0
120	0	10	0	2	0	0	0	0	1	0	0	1
160	0.693	10	1	2	0	0	0	0	1	0.693	0	1
218	1.099	10	2	2	0	0	0	0	1	1.099	0	1

4 Estimation and testing of fixed-effect panel-data systems

Given that (1) and (2) include fixed effects, the user must choose among FGLS (`xtgls`), OLS with panel-corrected standard errors (PCSE) (`xtpcse`), or fixed-effects regression (`areg` or `xtreg, fe`). This topic is explored at length by Beck and Katz (1995), and the key issues are neatly summarized by Wiggins (2001). However, only `xtgls` and `xtpcse` have all the options for the HPAC structure, so we will focus on those. See [XT] **xtgls** and [XT] **xtpcse** for details.

When `xtgls` is indicated, the command for HPAC SUR estimation is

```
. xtgls y c1 x1 d2-d7 c2 x2 d9-d14, i(pnl) t(week) p(c) c(psar1) nocon
```

When `xtpcse` is indicated, the commands for HPAC estimation are

```
. tsset pnl week
. xtpcse y c1 x1 d2-d7 c2 x2 d9-d14, c(psar1) nocon
```

Since our dataset has 14 panels and 9 time periods, FGLS is of course ruled out (Beck and Katz 1995, 637), so the output below (slightly edited) was produced with `xtpcse`:

```
Prais-Winsten regression, correlated panels corrected standard errors (PCSEs)

Group variable:   pnl                             Number of obs      =       126
Time variable:    week                            Number of groups   =        14
Panels:           correlated (balanced)           Obs per group: min =         9
Autocorrelation:  panel-specific AR(1)                           avg =         9
                                                                 max =         9
Estimated covariances      =       105            R-squared          =         X
Estimated autocorrelations =        14            Wald chi2(8)       =         X
Estimated coefficients     =        16            Prob > chi2        =         X

------------------------------------------------------------------------------
             |           Panel-corrected
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          c1 |   111.5311   6.208398    17.96   0.000    99.36284    123.6993
          x1 |   88.67886   3.943791    22.49   0.000    80.94918    96.40855
          d2 |   27.81103   1.400501    19.86   0.000     25.0661    30.55596
          d3 |   4.477261   3.969675     1.13   0.259   -3.303159    12.25768
          d4 |  15.006611   4.364777     3.45   0.001    6.511306     3.62092
          d5 |   -39.5333   3.116503   -12.69   0.000   -45.64154   -33.42507
          d6 |  -23.79004     2.0549   -11.58   0.000   -27.81757   -19.76251
          d7 |   9.415222    5.07866     1.85   0.064   -.5387687    19.36921
          c2 |   111.6787   8.215256    13.59   0.000    95.57709    127.7803
          x2 |   115.4025   5.231997    22.06   0.000     105.148     125.657
          d9 |   39.92659    3.15385    12.66   0.000    33.74516    46.10803
         d10 |  -22.60235    5.72391    -3.95   0.000     -33.821   -11.38369
         d11 |  -42.55708   5.499978    -7.74   0.000   -53.33684   -31.77732
         d12 |  -28.39879   5.906044    -4.81   0.000   -39.97442   -16.82315
         d13 |   29.44482   1.958789    15.03   0.000    25.60566    33.28397
         d14 |   78.11019   58.09246     1.34   0.179   -35.74894    191.9693
-------------+----------------------------------------------------------------
      rhos =    .249791   .0870025  .1838984  .3443183  -.2943893  -0.0062851
------------------------------------------------------------------------------
```

The R-squared and Wald statistics in the top section of the above output are obliterated for two reasons. First, they include the influences of the fixed-effect dummies, which serve only to control for the influences of unobserved variables ([R] **areg**, Stata 8, 87 or Stata 9, 95; Wooldridge 2003, 465–467). Second, the R-squared and Wald statistics in the `xtpcse` output are computed for the entire *system* of equations, which is meaningless for either equation individually. Fortunately, individual equations can be properly evaluated with `test`, as discussed below.

The coefficients and panel-corrected standard errors in the main body of the `xtpcse` output are correctly reported and consistent, but inefficient. The reported z statistics assume zero nulls and two-sided alternatives. See Baltagi (2001, 78) for a comparison of the properties of FGLS and OLS with PCSEs.

We can easily obtain Wald statistics for any linear combination of coefficients in the above output (Judge et al. 1988, 456–459). For any particular equation, we can use `test` to obtain the Wald statistic to determine the joint influence of all explanatory variables pertaining to that equation, but not including the influences of the fixed-effect dummies, as explained above. In (1), however, `x1` is the only explanatory variable for which we wish to know the effect on `y`, so the Wald statistic is simply the square of the z-score for `x1`, and the Wald statistic and the z-score are equivalent. Likewise, the z-score for `x2` provides the test for (2). Again, both `x1` and `x2` are tested against zero nulls and two-sided alternatives.

The test for a structural difference requires a joint test of all corresponding coefficients from the two equations in the system, in this case including the constant terms, since these too can vary by equation. For our equations, we thus require two linear restrictions to be jointly tested against two-sided alternatives:

```
. test (c1-c2 = 0) (x1-x2 = 0)
 ( 1)  c1 - c2 = 0
 ( 2)  x1 - x2 = 0
           chi2(  2) = 137.10
         Prob > chi2 =   0.0000
```

5 Acknowledgments

Sincere thanks to Brian Poi and Gustavo Sanchez for support in the development of the ideas presented here. Any remaining errors are, of course, mine alone.

6 References

Baltagi, B. H. 2001. *Econometric Analysis of Panel Data*. 2nd ed. New York: Wiley.

Beck, N. and J. N. Katz. 1995. What to do (and not to do) with time-series cross-section data. *American Political Science Review* 89: 634–647.

Judge, G. G., R. C. Hill, W. E. Griffiths, H. Lütkepohl, and T. C. Lee. 1988. *Introduction to the Theory and Practice of Econometrics*. 2nd ed. New York: Wiley.

Wiggins, V. 2001. FAQ: How does xtgls differ from regression clustered with robust standard errors?
http://www.stata.com/support/faqs/stat/xtgls_rob.html.

Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

—. 2003. *Introductory Econometrics: A Modern Approach*. 2nd ed. New York: South-Western.

About the Author

J. Lloyd Blackwell, III is Professor of Economics, Director of the Bureau of Business and Economic Research, and Director of the Master of Science in Applied Economics, University of North Dakota, Grand Forks, ND.