The Stata Journal

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

The Stata Journal (2006)
6, Number 4, pp. 482–496

Testing for cross-sectional dependence in panel-data models

Rafael E. De Hoyos
Development Prospects Group
The World Bank
Washington, DC
rdehoyos@worldbank.org

Vasilis Sarafidis
University of Sydney
Sydney, Australia
v.sarafidis@econ.usyd.edu.au

Abstract. This article describes a new Stata routine, `xtcsd`, to test for the presence of cross-sectional dependence in panels with many cross-sectional units and few time-series observations. The command executes three different testing procedures—namely, Friedman's (*Journal of the American Statistical Association* 32: 675–701) (FR) test statistic, the statistic proposed by Frees (*Journal of Econometrics* 69: 393–414), and the cross-sectional dependence (CD) test of Pesaran (*General diagnostic tests for cross-section dependence in panels* [University of Cambridge, Faculty of Economics, Cambridge Working Papers in Economics, Paper No. 0435]). We illustrate the command with an empirical example.

Keywords: st0113, xtcsd, panel data, cross-sectional dependence

1 Introduction

A growing body of the panel-data literature concludes that panel-data models are likely to exhibit substantial cross-sectional dependence in the errors, which may arise because of the presence of common shocks and unobserved components that ultimately become part of the error term, spatial dependence, and idiosyncratic pairwise dependence in the disturbances with no particular pattern of common components or spatial dependence. See, for example, Robertson and Symons (2000), Pesaran (2004), Anselin (2001), and Baltagi (2005, sec. 10.5). One reason for this result may be that during the last few decades we have experienced an ever-increasing economic and financial integration of countries and financial entities, which implies strong interdependencies between cross-sectional units. In microeconomic applications, the propensity of individuals to respond similarly to common "shocks", or common unobserved factors, may be plausibly explained by social norms, neighborhood effects, herd behavior, and genuinely interdependent preferences.

The impact of cross-sectional dependence in estimation naturally depends on a variety of factors, such as the magnitude of the correlations across cross sections and the nature of cross-sectional dependence itself. If we assume that cross-sectional dependence is caused by the presence of common factors, which are unobserved (and the effect of these components is therefore felt through the disturbance term) but uncorrelated with the included regressors, the standard fixed-effects (FE) and random-effects (RE) estimators are consistent, although not efficient, and the estimated standard errors are

biased. Thus different possibilities arise in estimation. For example, one may choose to retain the FE/RE estimators and correct the standard errors by following the approach proposed by Driscoll and Kraay (1998).[1] This method can be implemented in Stata by using the command `xtscc`, which is forthcoming to Statalist by Daniel Hoechle. Or, one may attempt to obtain an efficient estimator in the first place by using the methods put forward by Robertson and Symons (2000) and Coakley, Fuertes, and Smith (2002).

On the other hand, if the unobserved components that create interdependencies across cross sections are correlated with the included regressors, these approaches will not work and the FE and RE estimators will be biased and inconsistent. Here one may follow the approach proposed by Pesaran (2006). Another method would be to apply an instrumental variables (IV) approach using standard FE IV or RE IV estimators. However, in practice, finding instruments that are correlated with the regressors and not correlated with the unobserved factors would be difficult.

The impact of cross-sectional dependence in dynamic panel estimators is more severe. In particular, Phillips and Sul (2003) show that if there is sufficient cross-sectional dependence in the data and this is ignored in estimation (as it is commonly done by practitioners), the decrease in estimation efficiency can become so large that, in fact, the pooled (panel) least-squares estimator may provide little gain over the single-equation ordinary least squares. This result is important because it implies that if one decides to pool a population of cross sections that is homogeneous in the slope parameters but ignores cross-sectional dependence, then the efficiency gains that one had hoped to achieve, compared with running individual ordinary least-squares regressions for each cross section, may largely diminish.

Dealing specifically with short dynamic panel-data models, Sarafidis and Robertson (2006) show that if there is cross-sectional dependence in the disturbances, all estimation procedures that rely on IV and the generalized method of moments (GMM)—such as those by Anderson and Hsiao (1981), Arellano and Bond (1991), and Blundell and Bond (1998)—are inconsistent as N (the cross-sectional dimension) grows large, for fixed T (the panel's time dimension). This outcome is important given that error cross-section dependence is a likely practical situation and the desirable N-asymptotic properties of these estimators rely upon this assumption.[2]

The above indicates that testing for cross-sectional dependence is important in fitting panel-data models. When $T > N$, one may use for these purposes the Lagrange multiplier (LM) test, developed by Breusch and Pagan (1980), which is readily available in Stata through the command `xttest2` (Baum 2001, 2003, 2004). On the other hand, when $T < N$, the LM test statistic enjoys no desirable statistical properties in that it

1. Using cluster–robust standard errors will not help here because the correlations across groups of cross sections take nonzero values.

2. Intuitively, this result holds because for fixed T the common unobserved factor that is present in the disturbances is not averaged away to zero as $N \to \infty$, even if it is zero-mean distributed. Therefore, $p\lim_{N\to\infty} \left\{ \frac{1}{N} \sum_i^N (u_{it} u_{it-k}) \right\} \neq 0 \ \forall \ k$, which implies that there is no valid instrument to be used with respect to a lagged value of the dependent variable, regardless of how large the difference apart in time between the instrument and the endogenous regressor is. See Sarafidis and Robertson (2006, sec. 3) for more details.

exhibits substantial size distortions.[3] Thus there is clearly a need for testing for cross-sectional dependence in Stata when N is large and T is small—the most commonly encountered situation in panels.

This article describes a new Stata command that implements three different tests for cross-sectional dependence. The tests are valid when $T < N$ and can be used with balanced and unbalanced panels.

The rest of this article consists of the following: the next section describes three statistical procedures designed to test for cross-sectional dependence in large-N, small-T panels—namely, Pesaran's (2004) cross-sectional dependence (CD) test, Friedman's (1937) statistic, and the test statistic proposed by Frees (1995).[4] Section 3 describes the newly developed Stata command `xtcsd`. Section 4 illustrates using `xtcsd` by means of an empirical example based on gross product equations using a balanced panel dataset of states in the United States during 1970–1986. This is a widely cited dataset available from Baltagi's (2005) econometric textbook. A final section concludes the article.

2 Tests of cross-sectional dependence

Consider the standard panel-data model

$$y_{it} = \alpha_i + \beta' \mathbf{x}_{it} + u_{it}, \quad i = 1, \ldots, N \text{ and } t = 1, \ldots T \tag{1}$$

where $\mathbf{x}_{it}$ is a $K \times 1$ vector of regressors, β is a $K \times 1$ vector of parameters to be estimated, and α_i represents time-invariant individual nuisance parameters. Under the null hypothesis, u_{it} is assumed to be independent and identically distributed (i.i.d.) over periods and across cross-sectional units. Under the alternative, u_{it} may be correlated across cross sections, but the assumption of no serial correlation remains.

3. See Pesaran (2004) or Sarafidis, Yamagata, and Robertson (2006).

4. Two additional tests have been recently proposed by Sarafidis, Yamagata, and Robertson (2006) and Pesaran, Ullah, and Yamagata (2006). The SYR test is based on a Sargan's difference–type test and is relevant in short dynamic panel models. The PUY test is relevant in panel-data models with strictly exogenous regressors and normal errors. The SYR test involves computing Sargan's statistic for overidentifying restrictions based on two different GMM estimators: one that uses the full set of instruments available (including those with respect to lags of the dependent variable) and another that uses only a subset of instruments, in particular those with respect to the exogenous regressors. Under the null hypothesis of cross-sectional independence, both GMM estimators are consistent, whereas under the alternative of error cross-sectional dependence, the latter estimator remains consistent but the former does not. Hence, a large value of the difference between the two statistics would imply that the moment conditions with respect to lags of the dependent variable are not valid—a direct consequence of cross-sectional dependence. Since the proposed test can be implemented rather straightforwardly in Stata, the test is not discussed further here. For more details, see the reference above.
The PUY test statistic is essentially a bias-adjusted normal approximation to the LM test that is valid for N large and N small, in models with strictly exogenous regressors. Since the Pesaran et al. paper was made publicly available after the `xtcsd` command had been completed, we do not discuss this test any further.

Thus the hypothesis of interest is

$$H_0\text{: } \rho_{ij} = \rho_{ji} = \text{cor}\,(u_{it}, u_{jt}) = 0 \quad \text{for} \;\; i \neq j \tag{2}$$

versus

$$H_1\text{: } \rho_{ij} = \rho_{ji} \neq 0 \;\; \text{for some} \;\; i \neq j$$

where ρ_{ij} is the product-moment correlation coefficient of the disturbances and is given by

$$\rho_{ij} = \rho_{ji} = \frac{\sum_{t=1}^{T} u_{it} u_{jt}}{\left(\sum_{t=1}^{T} u_{it}^2\right)^{1/2} \left(\sum_{t=1}^{T} u_{jt}^2\right)^{1/2}}$$

The number of possible pairings (u_{it}, u_{jt}) rises with N.

2.1 Pesaran's CD test

In the context of seemingly unrelated regression estimation, Breusch and Pagan (1980) proposed an LM statistic, which is valid for fixed N as $T \to \infty$ and is given by

$$\text{LM} = T \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \widehat{\rho}_{ij}^2$$

where $\widehat{\rho}_{ij}$ is the sample estimate of the pairwise correlation of the residuals

$$\widehat{\rho}_{ij} = \widehat{\rho}_{ji} = \frac{\sum_{t=1}^{T} \widehat{u}_{it} \widehat{u}_{jt}}{\left(\sum_{t=1}^{T} \widehat{u}_{it}^2\right)^{1/2} \left(\sum_{t=1}^{T} \widehat{u}_{jt}^2\right)^{1/2}}$$

and $\widehat{u}_{it}$ is the estimate of u_{it} in (1). LM is asymptotically distributed as χ^2 with $N(N-1)/2$ degrees of freedom under the null hypothesis of interest. However, this test is likely to exhibit substantial size distortions when N is large and T is finite—a situation that is commonly encountered in empirical applications, primarily because the LM statistic is not correctly centered for finite T and the bias is likely to get worse with N large.

Pesaran (2004) has proposed the following alternative,

$$\text{CD} = \sqrt{\frac{2T}{N(N-1)}} \left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \widehat{\rho}_{ij} \right) \tag{3}$$

and showed that under the null hypothesis of no cross-sectional dependence $\text{CD} \xrightarrow{d} N\,(0,1)$ for $N \to \infty$ and T sufficiently large.

Unlike the LM statistic, the CD statistic has mean at exactly zero for fixed values of T and N, under a wide range of panel-data models, including homogeneous/heterogeneous

dynamic models and nonstationary models. For homogeneous and heterogeneous dynamic models, the standard FE and RE estimators are biased (see Nickell [1981] and Pesaran and Smith [1995]). However, the CD test is still valid because, despite the small-sample bias of the parameter estimates, the FE/RE residuals will have exactly mean zero even for fixed T, provided that the disturbances are symmetrically distributed.

For unbalanced panels, Pesaran (2004) proposes a slightly modified version of (3), which is given by

$$\text{CD} = \sqrt{\frac{2}{N(N-1)}} \left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \sqrt{T_{ij}} \widehat{\rho}_{ij} \right) \tag{4}$$

where $T_{ij} = \#\left(T_i \cap T_j\right)$ (i.e., the number of common time-series observations between units i and j),

$$\widehat{\rho}_{ij} = \widehat{\rho}_{ji} = \frac{\sum_{t \in T_i \cap T_j} \left(\widehat{u}_{it} - \overline{\widehat{u}}_i\right)\left(\widehat{u}_{jt} - \overline{\widehat{u}}_j\right)}{\left\{\sum_{t \in T_i \cap T_j} \left(\widehat{u}_{it} - \overline{\widehat{u}}_i\right)^2\right\}^{1/2} \left\{\sum_{t \in T_i \cap T_j} \left(\widehat{u}_{jt} - \overline{\widehat{u}}_j\right)^2\right\}^{1/2}}$$

and

$$\overline{\widehat{u}}_i = \frac{\sum_{t \in T_i \cap T_j} \widehat{u}_{it}}{\#\left(T_i \cap T_j\right)}$$

The modified statistic accounts for the fact that the residuals for subsets of t are not necessarily mean zero.

2.2 Friedman's test

Friedman (1937) proposed a nonparametric test based on Spearman's rank correlation coefficient. The coefficient can be thought of as the regular product-moment correlation coefficient, that is, in terms of proportion of variability accounted for, except that Spearman's rank correlation coefficient is computed from ranks. In particular, if we define $\{r_{i,1}, \ldots, r_{i,T}\}$ to be the ranks of $\{u_{i,1}, \ldots, u_{i,T}\}$ [such that the average rank is $(T + 1/2)$], Spearman's rank correlation coefficient equals[5]

$$r_{ij} = r_{ji} = \frac{\sum_{t=1}^{T} \{r_{i,t} - (T + 1/2)\}\{r_{j,t} - (T + 1/2)\}}{\sum_{t=1}^{T} \{r_{i,t} - (T + 1/2)\}^2}$$

Friedman's statistic is based on the average Spearman's correlation and is given by

$$R_{\text{ave}} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \widehat{r}_{ij}$$

5. Spearman's rank correlation coefficient as calculated by the Stata `spearman` command is slightly different in that it uses a definition of "average rank".

where $\widehat{r}_{ij}$ is the sample estimate of the rank correlation coefficient of the residuals. Large values of R_{ave} indicate the presence of nonzero cross-sectional correlations. Friedman showed that $\text{FR} = (T-1)\{(N-1)R_{\text{ave}}+1\}$ is asymptotically χ^2 distributed with $T-1$ degrees of freedom, for fixed T as N gets large. Originally Friedman devised the test statistic FR to determine the equality of treatment in a two-way analysis of variance.

The CD and R_{ave} share a common feature; both involve the sum of the pairwise correlation coefficients of the residual matrix rather than the sum of the squared correlations used in the LM test. This feature implies that these tests are likely to miss cases of cross-sectional dependence where the sign of the correlations is alternating—that is, where there are large positive and negative correlations in the residuals, which cancel each other out during averaging. Consider, for example, the following error structure of u_{it} under H_1,

$$u_{it} = \phi_i f_t + \varepsilon_{it} \tag{5}$$

where f_t represents the unobserved factor that generates cross-sectional dependence, ϕ_i indicates the impact of the factor on unit i, and ε_{it} is a pure idiosyncratic error with $f_t \sim \text{i.i.d.}\,(0,1)$, $\phi_i \sim \text{i.i.d.}\left(0,\sigma_\phi^2\right)$, and $\varepsilon_{it} \sim \text{i.i.d.}\left(0,\sigma_\varepsilon^2\right)$. Here we have

$$\text{cor}\,(u_{it},u_{jt}) = \frac{\text{cov}\,(u_{it},u_{jt})}{\sqrt{\text{var}\,(u_{it})}\sqrt{\text{var}\,(u_{jt})}} = \frac{E\,(\phi_i)\,E\,(\phi_j)}{\sqrt{E\,(u_{it}^2)}\sqrt{E\left(u_{jt}^2\right)}} = 0$$

and thereby the CD and R_{ave} statistics converge to 0 even if $f_t \neq 0$ and $\phi_i \neq 0$ for some i. This outcome implies that under alternative hypotheses of cross-sectional dependence in the disturbances with large positive and negative correlations but with $E\,(\phi_i) = 0$, these tests would lack power and therefore may not be reliable.

To see the relevance of the above argument, consider the initial panel-data model given by (1) and suppose that there is a single-factor structure in the disturbances, as in (5), except that the factor loadings are not mean zero, such that $E\,(\phi_i) \neq 0$. Apparently, the CD and R_{ave} tests would not be subject to the problem mentioned above in this case. However, there is a subtle thing that needs to be taken into account; in panels with N large and T finite, it is common practice to include common time effects (CTEs) in the regression model to capture "common trends" in the variation of the dependent variable across cross sections. Using CTEs is equivalent to time demeaning of the data, which implies that the initial panel-data model can now be written as

$$\begin{aligned}(y_{it}-\overline{y}_{.t}) &= (\alpha_i-\overline{\alpha})+\beta'\,(\mathbf{x}_{it}-\overline{\mathbf{x}}_{.t})+(u_{it}-\overline{u}_{.t})\\ (u_{it}-\overline{u}_{.t}) &= \left(\phi_i-\overline{\phi}\right)f_t+(\varepsilon_{it}-\overline{\varepsilon}_{.t})\end{aligned}$$

where $\overline{y}_{.t} = \frac{1}{N}\sum_i^N y_{it}$, and so on. As we can see, time demeaning of the data has transformed the disturbances in terms of deviations from time-specific averages, and therefore it has essentially removed the mean impact of the factors. This is the case unless of course the factor loadings are mean zero in the first place, in which case time demeaning is completely ineffective. Notice here two polar cases with regard to the variance of the factor loadings; at one extreme, if the variance of the ϕ_i's grows large,

time demeaning will be less effective because even if the mean impact of the factors has been removed, there is still a considerable amount of cross-sectional dependence left out in the disturbances. At the other extreme, if the variance of the ϕ_i's is zero, time demeaning removes cross-sectional dependence from the disturbances. Using CTEs will usually reduce cross-sectional dependence, but only to a certain extent.

Now suppose that the empirical researcher includes CTEs in the regression model and wants to see whether there is any cross-sectional dependence left out in the disturbances. Here $\operatorname{cov}\{(u_{it}-\overline{u}_{.t})(u_{jt}-\overline{u}_{.t})\} = E\left(\phi_i-\overline{\phi}\right)E\left(\phi_j-\overline{\phi}\right) = 0$. Thus the original problem emerges again in that the CD and R_{ave} tests will lack power to detect a false null hypothesis, even if there is plenty of cross-sectional dependence left out in the disturbances.[6]

2.3 Frees' test

Frees (1995, 2004) proposed a statistic that is not subject to this drawback.[7] In particular, the statistic is based on the sum of the squared rank correlation coefficients and equals

$$R^2_{\text{ave}} = \frac{2}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\widehat{r}^2_{ij}$$

As shown by Frees, a function of this statistic follows a joint distribution of two independently drawn χ^2 variables. In particular, Frees shows that

$$\begin{aligned}\text{FRE} &= N\left\{R^2_{\text{ave}}-(T-1)^{-1}\right\} \xrightarrow{d} Q = a(T)\left\{x^2_{1,T-1}-(T-1)\right\}\\ &\quad + b(T)\left\{x^2_{2,T(T-3)/2}-T(T-3)/2\right\}\end{aligned}$$

where $x^2_{1,T-1}$ and $x^2_{2,T(T-3)/2}$ are independently χ^2 random variables with $T-1$ and $T(T-3)/2$ degrees of freedom, respectively, $a(T) = 4(T+2)/\left\{5(T-1)^2(T+1)\right\}$ and $b(T) = 2(5T+6)/\{5T(T-1)(T+1)\}$. Thus the null hypothesis is rejected if $R^2_{\text{ave}} > (T-1)^{-1} + Q_q/N$, where Q_q is the appropriate quantile of the Q distribution.

6. Effectively, time demeaning causes the resulting factor loadings to be mean zero, which implies that the resulting correlation coefficients of the disturbances will alternate in sign, making the CD and R_{ave} tests inappropriate.

7. The testing procedure proposed by Sarafidis, Yamagata, and Robertson (2006) is not subject to this drawback either.

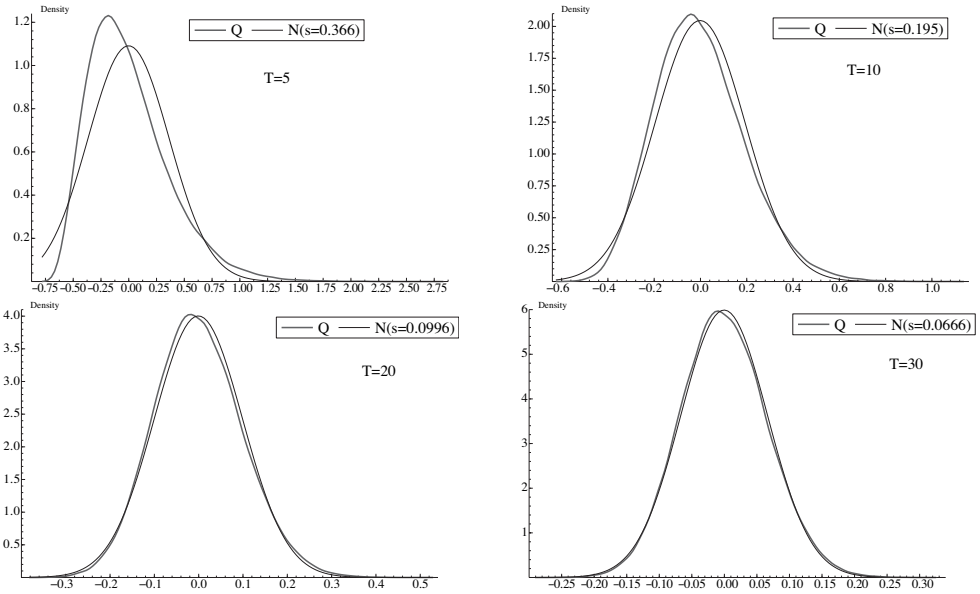

Figure 1: Normal approximation to the Q distribution (s denotes standard deviation)

The Q distribution is a (weighted) sum of two χ^2-distributed random variables and depends on the size of T. Hence, computation of the appropriate quantiles may be tedious. In cases where T is not small, Frees suggests using the normal approximation to the Q distribution by computing the variance of Q; i.e., we can use the following result,

$$\frac{\text{FRE}}{\sqrt{\text{Var}\,(Q)}} \approx N\,(0,1)$$

where

$$\text{Var}\,(Q) = \frac{32}{25}\frac{(T+2)^2}{(T-1)^3\,(T+1)^2} + \frac{4}{5}\frac{(5T+6)^2\,(T-3)}{T\,(T-1)^2\,(T+1)^2}$$

The accuracy of the normal approximation is illustrated in figure 1, which shows the density of Q for different values of T. As we can see, for small values of T the normal approximation to the Q distribution is poor. However, for T as large as 30, the approximation does well. Contrary to Pesaran's CD test, the tests by Frees and Friedman have been originally devised for static panels, and the finite-sample properties of the tests have not been investigated yet in dynamic panels.

3 The xtcsd command

The new Stata command `xtcsd` tests for the presence of cross-sectional dependence in FE and RE panel-data models. The command is suitable for cases where T is small as $N \to \infty$. It therefore complements the existing Breusch–Pagan LM test written by Christopher F. Baum, `xttest2`, which is valid for small N as $T \to \infty$. By making available a series of tests for cross-sectional dependence for cases where N is large and T is small, `xtcsd` closes an important gap in applied research.[8]

3.1 Syntax

`xtcsd` [`, pesaran friedman frees abs show`]

As with all other Stata cross-sectional time-series (`xt`) commands, the data need to be `tsset` before you use `xtcsd`. `xtcsd` is a postestimation command valid for use after running an FE or RE model.

3.2 Options

`pesaran` performs the CD test developed by Pesaran (2004) as explained in section 2.1. For balanced panels, `pesaran` estimates (3). For unbalanced panels, `pesaran` estimates (4). The CD statistic is normally distributed under the null hypothesis (2) for $T_i > k + 1$, and $T_{ij} > 2$ with sufficiently large N. Therefore, there must be enough cross-sectional units with common points in time to be able to implement the test.

`friedman` performs Friedman's test for cross-sectional dependence by using the nonparametric χ^2-distributed R_{ave} statistic (see section 2.2). For unbalanced panels, Friedman's test uses only the observations available for all cross-sectional units.

`frees` tests for cross-sectional dependence with Frees' Q distribution (T-asymptotically distributed). For unbalanced panels, Frees' test uses only the observations available for all cross-sectional units.[9] For $T > 30$, `frees` uses a normal approximation to obtain the critical values of the Q distribution.

8. `xtcsd` creates an $N \times N$ matrix of correlations of the residuals. Hence, the maximum number of cross-sectional units that can be handled by `xtcsd` will be bounded by the matrix size capabilities of the version of Stata being used (see `help limits`). If N is prohibitively large, one can run `xtcsd` for different subsets of the sample. Rejecting the null hypothesis in all subsets would serve as an indication that there is cross-sectional dependence in the disturbances that needs to be taken into account.

9. This condition could be highly restrictive when only a few cross-sectional units show many missing values. In such cases, it might be preferable to drop the *problematic* cross-sectional units—i.e., those with many missing values—and perform the test using only the cross-sectional units with a relatively large number of observations.

`abs` computes the average absolute value of the off-diagonal elements of the cross-sectional correlation matrix of residuals. This option is useful to identify cases of cross-sectional dependence where the sign of the correlations is alternating, with the likely result of making the `pesaran` and `friedman` tests unreliable (see section 2.2).

`show` shows the cross-sectional correlation matrix of residuals.

4 Application

We illustrate `xtcsd` with an empirical example taken from Baltagi (2005, 25). The example refers to a Cobb–Douglas production function relationship investigating the productivity of public capital in private production. The dataset consists of a balanced panel of 48 U.S. states, each observed over 17 years (1970–1986). This dataset and some explanatory notes can be found on the Wiley web site.[10]

Following Munnell (1990) and Baltagi and Pinnoi (1995), Baltagi (2005) considers the following relationship,

$$\ln \texttt{gsp}_{it} = \alpha + \beta_1 \ln \texttt{p\_cap}_{it} + \beta_2 \ln \texttt{pc}_{it} + \beta_3 \ln \texttt{emp}_{it} + \beta_4 \texttt{unemp}_{it} + u_{it} \qquad (6)$$

where $\texttt{gsp}_{it}$ denotes gross product in state i at time t; `p_cap` denotes public capital including highways and streets, water and sewer facilities, and other public buildings; `pc` denotes the stock of private capital; `emp` is labor input measured as employment in nonagricultural payrolls; and `unemp` is the state unemployment rate included to capture business cycle effects.

We begin the exercise by downloading the data and declaring that it has a panel-data format:

```
. use http://www.econ.cam.ac.uk/phd/red29/xtcsd_baltagi.dta
. tsset id t
       panel variable:  id (strongly balanced)
        time variable:  t, 1970 to 1986
```

Once the dataset is ready for undertaking panel-data analysis, we run a version of (6) where we assume that u_{it} is formed by a combination of a fixed component specific to the state and a random component that captures pure noise. Below are the results of the model using the FE estimator, also reported in Baltagi (2005, 26):

10. The database in plain format is available from http://www.wiley.com/legacy/wileychi/baltagi/supp/PRODUC.prn; in the Stata Command window, type `net from http://www.econ.cam.ac.uk/phd/red29/` to get the data in Stata format.

```
. xtreg lngsp lnpcap lnpc lnemp unemp, fe
Fixed-effects (within) regression               Number of obs      =       816
Group variable (i): id                          Number of groups   =        48

R-sq:  within  = 0.9413                         Obs per group: min =        17
       between = 0.9921                                        avg =      17.0
       overall = 0.9910                                        max =        17

                                                F(4,764)           =   3064.81
corr(u_i, Xb)  = 0.0608                         Prob > F           =    0.0000

------------------------------------------------------------------------------
       lngsp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lnpcap |  -.0261493   .0290016    -0.90   0.368    -.0830815    .0307829
        lnpc |   .2920067   .0251197    11.62   0.000     .2426949    .3413185
       lnemp |   .7681595   .0300917    25.53   0.000     .7090872    .8272318
       unemp |  -.0052977   .0009887    -5.36   0.000    -.0072387   -.0033568
       _cons |   2.352898   .1748131    13.46   0.000     2.009727    2.696069
-------------+----------------------------------------------------------------
     sigma_u |  .09057293
     sigma_e |  .03813705
         rho |   .8494045   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(47, 764) =    75.82             Prob > F = 0.0000
```

According to the results, once we account for state FE, public capital has no effect upon state gross product in the United States. An assumption implicit in estimating (6) is that the cross-sectional units are independent. The `xtcsd` command allows us to test the following hypothesis:

H_0: cross-sectional independence

To test this hypothesis, we use the `xtcsd` command after fitting the above panel-data model. We initially use Pesaran's (2004) CD test:

```
. xtcsd, pesaran abs

Pesaran's test of cross sectional independence =    30.368, Pr = 0.0000

Average absolute value of the off-diagonal elements =     0.442
```

As we can see, the CD test strongly rejects the null hypothesis of no cross-sectional dependence. Although it is not the case here, a possible drawback of the CD test is that adding up positive and negative correlations may result in failing to reject the null hypothesis even if there is plenty of cross-sectional dependence in the errors. Including the `abs` option in the `xtcsd` command, we can get the average *absolute* correlation of the residuals. Here the average absolute correlation is 0.442, which is a very high value. Hence, there is enough evidence suggesting the presence of cross-sectional dependence in (6) under an FE specification.

Next we corroborate these results by using the remaining two tests explained in section 2, i.e., Frees (1995) and Friedman (1937):

```
. xtcsd, frees

  Frees' test of cross sectional independence =     8.386
|-------------------------------------------------------------|
  Critical values from Frees' Q distribution
                     alpha = 0.10 :   0.1521
                     alpha = 0.05 :   0.1996
                     alpha = 0.01 :   0.2928
. xtcsd, friedman

Friedman's test of cross sectional independence =   152.804, Pr = 0.0000
```

As we would have expected from the highly significant results of the CD test, both Frees' and Friedman's tests reject the null of cross-sectional independence. Since $T \leq 30$, Frees' test provides the critical values for $\alpha = 0.10$, $\alpha = 0.05$, and $\alpha = 0.01$ from the Q distribution. Frees' statistic is larger than the critical value with at least $\alpha = 0.01$.

Baltagi also reports the results of the model using the RE estimator. The results are shown below:

```
. xtreg lngsp lnpcap lnpc lnemp unemp, re
Random-effects GLS regression                   Number of obs      =       816
Group variable (i): id                          Number of groups   =        48

R-sq:  within  = 0.9412                         Obs per group: min =        17
       between = 0.9928                                        avg =      17.0
       overall = 0.9917                                        max =        17

Random effects u_i ~ Gaussian                   Wald chi2(4)       =  19131.09
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000
```

lngsp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnpcap	.0044388	.0234173	0.19	0.850	-.0414583	.0503359
lnpc	.3105483	.0198047	15.68	0.000	.2717317	.3493649
lnemp	.7296705	.0249202	29.28	0.000	.6808278	.7785132
unemp	-.0061725	.0009073	-6.80	0.000	-.0079507	-.0043942
_cons	2.135411	.1334615	16.00	0.000	1.873831	2.39699
sigma_u	.0826905					
sigma_e	.03813705					
rho	.82460109	(fraction of variance due to u_i)				

The results of this second model are in line with those of the previous one, with public capital having no significant effects upon gross state output. We now test for cross-sectional independence by using the new RE specification:

(*Continued on next page*)

```
. xtcsd, pesaran

Pesaran's test of cross sectional independence =    29.079, Pr = 0.0000
. xtcsd, frees

  Frees' test of cross sectional independence =     8.298
|-------------------------------------------------------------|
  Critical values from Frees' Q distribution
                     alpha = 0.10 :   0.1521
                     alpha = 0.05 :   0.1996
                     alpha = 0.01 :   0.2928
. xtcsd, friedman

Friedman's test of cross sectional independence =   144.941, Pr = 0.0000
```

The conclusion with respect to the existence or not of cross-sectional dependence in the errors is not altered. The results show that there is enough evidence to reject the null hypothesis of cross-sectional independence. The newly developed `xtcsd` Stata command shows an easy way of performing three popular tests for cross-sectional dependence.

5 Concluding remarks

This article has described a new Stata postestimation command, `xtcsd`, which tests for the presence of cross-sectional dependence in FE and RE panel-data models. The command executes three different testing procedures—namely, Friedman's (1937) test statistic, the statistic proposed by Frees (1995), and the CD test developed by Pesaran (2004). These procedures are valid when T is fixed and N is large.[11] `xtcsd` can also perform Pesaran's CD test for unbalanced panels.

Our view is that all these tests for cross-sectional dependence should not be regarded as competing but rather as complementary. If T is large relative to N, the LM test may be used. If N is large relative to T and the model is static, all different tests provided by `xtcsd` may be suitable, unless the empirical researcher has reason to believe that the correlation coefficients of the disturbances alternate in sign (or common time effects have been included in the model). In that case only the Frees test may be used.[12] One can ascertain whether this is the case by using the option `abs`, which computes the average absolute value of the off-diagonal elements of the cross-sectional correlation matrix of the residuals. If this takes a large value and the different tests provide contradicting results in the sense that Pesaran's and Friedman's tests fail to reject the null hypothesis, whereas Frees' test does not, inferences should be based on the latter. In dynamic panels, Pesaran's test remains valid under FE/RE estimation (even if the estimated parameters are biased) and therefore it may be the preferred choice, since the properties of the remaining tests in dynamic panels are not yet known. On the other hand, if common time effects have been included in the dynamic panel (and the panel is short), the test by Sarafidis, Yamagata, and Robertson (2006) may be used.

11. The CD test may also be used with both T and N large.
12. However, Pesaran, Ullah, and Yamagata (2006) indicate that Frees' test may not work well in models with explanatory variables when N is large.

In conclusion, the `xtcsd` command complements the Stata command `xttest2` that tests for the presence of error cross-sectional dependence with T large and finite N. Hence, `xtcsd` closes an important gap in applied research.

6 Acknowledgments

Our code benefited greatly from Christopher F. Baum's `xttest2`. We thank David Drukker and an anonymous referee for useful suggestions.

7 References

Anderson, T. W., and C. Hsiao. 1981. Estimation of dynamic models with error components. *Journal of the American Statistical Association* 76: 598–606.

Anselin, L. 2001. Spatial Econometrics. In *A Companion to Theoretical Econometrics*, ed. B. H. Baltagi, 310–330. Oxford: Blackwell Scientific Publications.

Arellano, M., and S. Bond. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies* 58: 277–297.

Baltagi, B. H. 2005. *Econometric Analysis of Panel Data*. 3rd ed. New York: Wiley.

Baltagi, B. H., and N. Pinnoi. 1995. Public capital stock and state productivity growth: Further evidence from an error components model. *Empirical Economics* 20: 351–359.

Baum, C. F. 2001. Residual diagnostics for cross-section time-series regression models. *Stata Journal* 1: 101–104.

———. 2003. Software updates: Residual diagnostics for cross-section time-series regression models. *Stata Journal* 3: 211.

———. 2004. Software updates: Residual diagnostics for cross-section time-series regression models. *Stata Journal* 4: 224.

Blundell, R., and S. Bond. 1998. Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87: 115–143.

Breusch, T., and A. Pagan. 1980. The Lagrange multiplier test and its application to model specification in econometrics. *Review of Economic Studies* 47: 239–253.

Coakley, J., A. Fuertes, and R. Smith. 2002. A principal components approach to cross-section dependence in panels. Unpublished manuscript.

Driscoll, J., and A. C. Kraay. 1998. Consistent covariance matrix estimation with spatially dependent data. *Review of Economics and Statistics* 80: 549–560.

Frees, E. W. 1995. Assessing cross-sectional correlation in panel data. *Journal of Econometrics* 69: 393–414.

———. 2004. *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge: Cambridge University Press.

Friedman, M. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* 32: 675–701.

Munnell, A. 1990. Why has productivity growth declined? Productivity and public investment. *New England Economic Review* (January/February): 3–22.

Nickell, S. J. 1981. Biases in dynamic models with fixed effects. *Econometrica* 49: 1417–1426.

Pesaran, M. H. 2004. General diagnostic tests for cross section dependence in panels. University of Cambridge, Faculty of Economics, Cambridge Working Papers in Economics No. 0435.

———. 2006. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrics* 74: 967–1012.

Pesaran, M. H., and R. Smith. 1995. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics* 68: 79–113.

Pesaran, M. H., A. Ullah, and T. Yamagata. 2006. A bias-adjusted test of error cross section dependence. http://www.econ.cam.ac.uk/faculty/pesaran/PUY10May06.pdf.

Phillips, P., and D. Sul. 2003. Dynamic panel estimation and homogeneity testing under cross section dependence. *Econometrics Journal* 6: 217–259.

Robertson, D., and J. Symons. 2000. Factor residuals in SUR regressions: Estimating panels allowing for cross sectional correlation. Unpublished manuscript.

Sarafidis, V., and D. Robertson. 2006. On the impact of cross section dependence in short dynamic panel estimation.
http://www.econ.cam.ac.uk/faculty/robertson/csd.pdf.

Sarafidis, V., T. Yamagata, and D. Robertson. 2006. A test of cross section dependence for a linear dynamic panel model with regressors.
http://www.econ.cam.ac.uk/faculty/robertson/HCSDtest14Feb06.pdf.

About the authors

Rafael E. De Hoyos works as a researcher at the Development Economics Prospects Group, the World Bank. His research includes topics such as policy evaluation, microeconometrics, and the economics of poverty and inequality.

Vasilis Sarafidis is a lecturer at the University of Sydney, Discipline of Econometrics and Business Statistics. His current research interests focus on GMM estimation of linear dynamic panel-data models with error cross-section dependence.