Journal of Statistical Software

January 2017, Volume 76, Issue 6. doi: 10.18637/jss.v076.i06

A Panel Data Toolbox for MATLAB

Inmaculada C. Álvarez
Universidad Autónoma
de Madrid

Javier Barbero
Universidad Autónoma
de Madrid

José L. Zofío
Universidad Autónoma
de Madrid

Abstract

Panel Data Toolbox is a new package for MATLAB that includes functions to estimate the main econometric methods of balanced and unbalanced panel data analysis. The package includes code for the standard fixed, between and random effects estimation methods, as well as for the existing instrumental panels and a wide array of spatial panels. A full set of relevant tests is also included. This paper describes the methodology and implementation of the functions and illustrates their use with well-known examples. We perform numerical checks against other popular commercial and free software to show the validity of the results.

Keywords: panel data, instrumental panel, spatial panel, econometrics, MATLAB.

1. Introduction

Panel data econometrics has grown in importance over the past decades due to increase in the availability of data related to units that are observed over a long period of time. Panel data econometric methods are available in Stata (StataCorp 2015) and R (R Core Team 2016), but there is a lack of a full set of functions for MATLAB (The MathWorks Inc. 2015).

The **Panel Data Toolbox** introduces such set of functions, including estimation methods for the standard fixed, between and random effects models, both balanced and unbalanced, as well as instrumental panel data models, including the error components by Baltagi (1981), and, finally, recently introduced spatial panels, (Kapoor, Kelejian, and Prucha 2007; Baltagi and Liu 2011). Numerical checks against Stata and R using well-known classical examples show that the estimated coefficients and t statistics are consistent with those obtained with the new MATLAB toolbox.[1]

[1] This paper corresponds to version 2.0 of the **Panel Data Toolbox** released in June 2015. The change log from the previous version, dating back to October 2013, can be found on `http://www.paneldatatoolbox.com/`.

A full set of corresponding tests is included for poolability of the data, individual effects, fixed and random effects, serial correlation, and cross-sectional dependence. An overidentification test is also available for instrumental panels, as well as tests for spatial autocorrelation.

Spatial econometrics in MATLAB can be estimated using the **Econometrics Toolbox** (LeSage and Pace 2009), which uses maximum likelihood and Bayesian methods, and using maximum likelihood methods (Elhorst 2014a). In this new **Panel Data Toolbox** we use a generalized spatial two stage least squares (GS2SLS) estimator for spatial panels following Kapoor *et al.* (2007) and Baltagi and Liu (2011).

Panel Data Toolbox is available as free software, under the GNU General Public License version 3, and can be downloaded from `http://www.paneldatatoolbox.com/`, with all the supplementary material (data, examples and source code) to replicate all the results presented in this paper. The toolbox is also hosted on an open source repository on GitHub at `https://github.com/javierbarbero/PanelDataMATLAB`.

The paper is organized as follows. Section 3 presents the estimation methods for panel data models. Testing procedures are shown in Section 4. Numerical checks against Stata and R are presented in Section 5. Section 6 concludes.

2. Data structures

Panel data contains units (individuals, firms, countries, etc.) that are observed over a long period of time. Units are usually denoted by $i = 1, 2, \dots, n$, and T_i is the number of time periods for which unit i is observed. This toolbox handles both balanced and unbalanced panel data, without any previous sorting required, as the toolbox orders the data internally. The total number of observations is $N = \sum_{i=1}^{n} T_i$, and simplifies to $N = nT$ in case of a balanced panel where $T_i = T \ \forall i$.

Data are managed as regular MATLAB vectors and matrices, constituting the inputs of the estimation functions. All estimation functions return a structure `estout` that contains fields with the estimation results as well as the input of the estimation function. Fields can be accessed directly using the dot notation and the whole structure can be used as an input to other functions that print results (e.g., `estdisp`) or perform postestimation tests.

Some of the fields of the `estout` structure are the following:[2]

- `y` and `X`: Contain the dependent and the independent variables, respectively.
- `n`, `T` and `N`: Number of entities, time periods, and total number of observations.
- `k` and `l`: Number of explanatory variables and instruments.
- `coef`, `varcoef` and `stderr`: Estimated coefficients, estimated covariance matrix, and estimated standard errors.
- `yhat` and `res`: Fitted values and residuals.

Testing functions take as input a `estout` structure and return as output a `testout` structure with the results of the test. The common fields of the `testout` structure are the following:

[2] For a full list see the help of the function typing `help estout` in MATLAB.

- `test`: Name of the test performed.
- `value`: Value/score of the test.
- `df`: Degrees of freedom.
- `p`: Associated p value.

3. Model estimation

The starting formulation is the panel data model with specific individual effects:

$$y_{it} = \alpha + X_{it}\beta + \mu_i + v_{it}, \qquad i = 1, \dots, n, \quad t = 1, \dots, T_i. \tag{1}$$

where μ_i represents the ith invariant time individual effect and v_{it} the disturbance, with $v_{it} \sim$ i.i.d$(0, \theta_v^2)$, $\mathsf{E}(v_i) = 0$, $\mathsf{E}(v_i v_i^\top) = \theta_v^2 I_T$ and $\mathsf{E}(v_i v_j) = 0$ for $i \neq j$, with I_T the $T \times T$ identity matrix.

3.1. Basic panel models

As a classic application we use the Munnell (1990) and Baltagi (2008) data. Munnell (1990) suggests a Cobb-Douglas production function using data for 48 U.S. states over 17 periods (1970–1986). The dependent variable, output of the production function, is the gross state product, `log(gsp)`, and the explanatory ones are public capital, `log(pcap)`, private capital, `log(pc)`, employment, `log(emp)`, and the unemployment rate, `unemp`.[3]

```
load('MunnellData')
y = log(gsp);
X = [log(pcap), log(pc), log(emp), unemp];
ynames = {'lgsp'};
xnames = {'lpcap', 'lpc', 'lemp', 'unemp'};
```

We create a vector `y` containing the dependent variable and a matrix `X` with the explanatory variables. A vector of ones for the constant term should not be added to `X` because it is included internally by the estimation functions. The variables `ynames` and `xnames` are cell arrays of strings that contain the names of the variables that are subsequently used when displaying the results of the estimation.

Panel data models are estimated using the `panel(id, time, y, X, method, options)` function, where `id` and `time` are vectors of unit and time indexes, `y` is the vector of the dependent variable, `X` is the matrix of explanatory variables, and `method` is a string that specifies the panel data estimation method to be used among the following:

- `po`: For a pooling estimation.
- `fe`: For a fixed effects (within) estimation.
- `be`: For a between estimation.

[3] The Munnell (1990) data are available in MATLAB format in the supplementary file `MunnellData.mat`.

- `re`: For a random effects GLS estimation.

These estimation methods are explained in the following sections. `options` is an optional list of parameter-value pairs to specify advanced estimating options.

Fixed effects

Under typical specifications, individual effects are correlated with the explanatory variables: $\mathsf{COV}(X_{it}, \mu_i) \neq 0$, which motivates the use of the fixed-effects (within) estimation, so as to capture unobserved heterogeneity (Baltagi 2008).

In this context, including individual effects on the error component while performing OLS (ordinary least squares) results into a biased estimation. In order to extract these effects, the within estimator of the parameters is computed using OLS:

$$\hat{\beta}_{fe} = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{y}, \tag{2}$$

where $\tilde{y} = y - \bar{y}$ and $\tilde{X} = X - \bar{X}$ are the transformed variables in deviations from the group means, $\bar{y}$ and $\bar{X}$. It is called "within" estimator because it takes into account the variations in each group. This estimator is unbiased and consistent for $n \to \infty$. Statistical inference is generally based on the asymptotic variance-covariance matrix:

$$\mathsf{VAR}(\hat{\beta}_{fe}) = S^2 (\tilde{X}^\top \tilde{X})^{-1}, \tag{3}$$

where S^2 denotes the residual variance: $S^2 = (e^\top e)/(N - n - k)$, with residuals $e = \tilde{y} - (\tilde{X}\hat{\beta}_{fe})$. Finally, inference can be performed using the standard t and F tests.

The `panel` function implements the estimation of fixed effects panel data models in MATLAB:

```
fe = panel(id, year, y, X, 'fe');
fe.ynames = ynames;
fe.xnames = xnames;
estdisp(fe);
```

```
Panel: Fixed effects (within) (FE)

N = 816  n = 48  T = 17 (Balanced panel)
R-squared = 0.94134    Adj R-squared = 0.93742
Wald F(4, 764) = 3064.808435 p-value = 0.0000
RSS = 1.111189 ESS = 90964.408970 TSS = 90964.408970

----------------------------------------------------------------------
           lgsp |   Coefficient    Std. Error      t-stat    p-value
----------------------------------------------------------------------
          lpcap |     -0.026150      0.029002     -0.9017    0.368
            lpc |      0.292007      0.025120     11.6246    0.000 ***
           lemp |      0.768159      0.030092     25.5273    0.000 ***
          unemp |     -0.005298      0.000989     -5.3582    0.000 ***
----------------------------------------------------------------------
```

The function `estdisp` is used to display the estimation results taking the names of the variables specified in the fields `ynames` and `xnames` of the `estout` structure that is returned from the `panel` function.[4]

The individual effects, with their standard errors and significance test, can be recovered with the `ieffects` command, and conveniently displayed with the `ieffectsdisp` function. They are computed as follows:

$$\hat{\mu} = \bar{y} - \bar{X}\beta, \tag{4}$$

$$\mathsf{VAR}(\mu_i) = \frac{\tilde{\sigma}_v^2}{T_i} + \bar{X}\mathsf{VAR}(\hat{\beta})\bar{X}^\top. \tag{5}$$

```
ieff = ieffects(fe);
ieffectsdisp(fe);

Individual Effects
----------------------------------------------------------------
     id  |       ieffect    Std. Error      t-stat    p-value
----------------------------------------------------------------
       1 |      2.201617      0.176004     12.5089    0.000 ***
       2 |      2.368088      0.175188     13.5174    0.000 ***
       3 |      2.263016      0.167172     13.5371    0.000 ***
       4 |      2.500423      0.201219     12.4264    0.000 ***
*** output cropped to save space ***
      45 |      2.446782      0.188093     13.0083    0.000 ***
      46 |      2.293150      0.171526     13.3691    0.000 ***
      47 |      2.328960      0.179153     12.9998    0.000 ***
      48 |      2.648557      0.178920     14.8030    0.000 ***
----------------------------------------------------------------
```

An "overall constant term", computed as the mean of the individual effects, can be calculated and displayed adding the parameter `'overall'` to the `ieffects` or `ieffectsdisp` functions.

```
ieffOver = ieffects(fe, 'overall');
ieffectsdisp(fe, 'overall');

Individual Effects
----------------------------------------------------------------
     id  |       ieffect    Std. Error      t-stat    p-value
----------------------------------------------------------------
 OVERALL |      2.352899      0.174808     13.4599    0.000 ***
----------------------------------------------------------------
```

Between estimation

[4] If variables `y` and `x` are in the `table` format introduced in MATLAB R2013b, the names of those variables are automatically assigned to the `ynames` and `xnames` fields when calling the estimation function.

The between estimation is performed by applying OLS to the transformed variables:

$$\hat{\beta}_{be} = (\bar{X}^\top \bar{X})^{-1} \bar{X}^\top \bar{y}, \tag{6}$$

where $\bar{y}$ and $\bar{X}$ are the group means of the variables. It is called "between" estimator because it takes into account the variation between groups. Again, statistical inference is based on the asymptotic variance-covariance matrix:

$$\mathsf{VAR}(\hat{\beta}_{be}) = S^2 (\bar{X}^\top \bar{X})^{-1}, \tag{7}$$

where S^2 denotes the residual variance: $S^2 = (e^\top e)/(n-k)$, with residuals $e = \bar{y} - \bar{X}\hat{\beta}_{be}$. The `panel` function implements the between estimation in MATLAB:

```
be = panel(id, year, y, X, 'be');
be.ynames = ynames;
be.xnames = xnames;
estdisp(be);
```

```
Panel: Between estimation (BE)

N = 816  n = 48  T = 17 (Balanced panel)
R-squared = 0.99391    Adj R-squared = 0.99334
Wald F(4, 43) = 1754.114154 p-value = 0.0000
RSS = 0.297701 ESS = 90965.222458 TSS = 90965.222458

-----------------------------------------------------------------------
           lgsp |   Coefficient    Std. Error      t-stat    p-value
-----------------------------------------------------------------------
          lpcap |      0.179365      0.071972      2.4922    0.017 **
            lpc |      0.301954      0.041821      7.2201    0.000 ***
           lemp |      0.576127      0.056375     10.2196    0.000 ***
          unemp |     -0.003890      0.009908     -0.3926    0.697
          CONST |      1.589444      0.232980      6.8222    0.000 ***
-----------------------------------------------------------------------
```

Random effects model

In the panel data model (1) the loss of degrees of freedom can be avoided if the individual effects can be assumed random, where the error component $u_{it} = \mu_i + v_{it}$ includes the ith invariant time individual effects μ_i and the disturbance v_{it}.

$$y_{it} = \alpha + X_{it}\beta + u_{it}, \qquad i = 1, \ldots, n, \quad t = 1, \ldots, T_i. \tag{8}$$

The individual effect μ_i is assumed independent of the disturbance v_{it}. In addition, individual effects and disturbances are independent of the explanatory variables; i.e., $\mathsf{COV}(X_{it}, \mu_i) = 0$ and $\mathsf{COV}(X_{it}, v_{it}) = 0$ for all i and t. For this reason, the random effects model is an appropriate specification in the analysis of n individuals randomly drawn from a large population.

In this context, n is usually large and a fixed effects model would lead to a loss of degrees of freedom.

Following the formalization of Wallace and Hussain (1969), as stated in Baltagi (2008), the composed error component has the following properties:

$$\mathsf{E}(\mu_i) = \mathsf{E}(v_{it}) = \mathsf{E}(\mu_i v_{it}) = 0, \tag{9}$$

$$\mathsf{E}(\mu_i \mu_j) = \begin{cases} \sigma_\mu^2 & i \neq j \\ 0 & i = j \end{cases} \qquad \mathsf{E}(v_i v_j) = \begin{cases} \sigma_v^2 & i \neq j \\ 0 & i = j. \end{cases} \tag{10}$$

This results in a block-diagonal covariance matrix with serial correlation over time, only between disturbances of the same individual and zero otherwise:

$$\mathsf{COV}(u_{it}, u_{js}) = \begin{cases} \sigma_\mu^2 + \sigma_v^2 & i = j, t = s \\ \sigma_\mu^2 & i = j, t \neq s. \end{cases} \tag{11}$$

This implies the following correlation coefficient between disturbances:

$$\rho = \mathsf{CORR}(u_{it}, u_{js}) = \begin{cases} 1 & i = j, t = s \\ \sigma_\mu^2 / (\sigma_\mu^2 + \sigma_v^2) & i = j, t \neq s. \end{cases} \tag{12}$$

Therefore, the covariance matrix can be computed as follows:

$$\Omega = \mathsf{E}(uu^\top) = \sigma_\mu^2 (I_n \otimes J_T) + \sigma_v^2 (I_n \otimes I_T), \tag{13}$$

where J_T is a matrix of ones of size T and the homoscedastic variance is $\mathsf{VAR}(u_{it}) = \sigma_\mu^2 + \sigma_v^2$ for all i and t. In this case, the GLS (generalized least squares) method yields an efficient estimator of the parameters,

$$\hat{\beta}_{re} = (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} y, \tag{14}$$

with $\Omega^{-1} = 1/\sigma_1^2 P + 1/\sigma_v^2 Q$, where $\sigma_1^2 = T\sigma_\mu^2 + \sigma_v^2$, and P and Q are the matrices that compute the group means and the differences with respect to the group means, respectively. In order to obtain the GLS estimator of the regression coefficients, it is necessary to estimate the Ω^{-1} matrix of dimension $nT \times nT$. Fuller and Battese (1973, 1974) suggest premultiplying the model by $\sigma_v \Omega^{-1/2}$, which is equivalent to computing a quasi-time demeaning of the variables $\tilde{y}_{it} = y_{it} - \theta_i \bar{y}_i$ and $\tilde{X}_{it} = X_{it} - \theta_i \bar{X}_i$, where

$$\theta_i = 1 - \sqrt{\frac{\sigma_v^2}{T_i \sigma_\mu^2 + \sigma_v^2}}. \tag{15}$$

Then, the random effects GLS estimation is computed as

$$\hat{\beta}_{re} = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top y. \tag{16}$$

Now the question is how to obtain estimates of σ_v^2, σ_μ^2 and σ_1^2. Among the different methods proposed in the literature, Swamy and Arora (1972) suggest using the within regression residuals to compute $\hat{\sigma}_v^2$ and the residuals from the between regression to compute $\hat{\sigma}_1^2$. From these estimates $\hat{\sigma}_\mu^2$ is calculated as:[5]

$$\hat{\sigma}_\mu^2 = \sigma_1^2 - \frac{\sigma_v^2}{\bar{T}}, \tag{17}$$

where $\bar{T}$ is the harmonic mean of T in case of an unbalanced panel, and simple T if the panel is balanced. The random effects estimator (16) is a weighted average of the within and the between estimators. In this case, the asymptotic variance-covariance matrix for statistical inference is:

$$\mathsf{VAR}(\hat{\beta}_{re}) = S^2(\tilde{X}^\top \tilde{X})^{-1}, \tag{18}$$

where, once again, S^2 denotes the residual variance: $S^2 = (e^\top e)/(N - k)$, with residuals $e = \tilde{y} - \tilde{X}\hat{\beta}_{re}$.

The `panel` function implements the estimation of random effects panel data in MATLAB:

```
re = panel(id, year, y, X, 're');
re.ynames = ynames;
re.xnames = xnames;
estdisp(re);

Panel: Random effects (RE)

N = 816  n = 48  T = 17 (Balanced panel)
R-squared = 0.99167    Adj R-squared = 0.99163
Wald Chi2(4) = 19131.085009 p-value = 0.0000
RSS = 1.187864 ESS = 90964.332295 TSS = 90964.332295

-----------------------------------------------------------------------
           lgsp |   Coefficient    Std. Error      z-stat    p-value
-----------------------------------------------------------------------
          lpcap |      0.004439      0.023417      0.1895    0.850
            lpc |      0.310548      0.019805     15.6805    0.000 ***
           lemp |      0.729671      0.024920     29.2803    0.000 ***
          unemp |     -0.006172      0.000907     -6.8033    0.000 ***
          CONST |      2.135411      0.133461     16.0002    0.000 ***
-----------------------------------------------------------------------
sigma_mu = 0.082691      rho_mu = 0.824601
 sigma_v = 0.038137     sigma_1 = 0.083206
   theta = 0.888835
-----------------------------------------------------------------------
```

The estimation output displays the estimated $\hat{\sigma}_\mu$, $\hat{\sigma}_v$, $\hat{\sigma}_1$, and $\hat{\theta}$, as well as `rho_mu`, which is the fraction of variance due to the individual effects computed as $\hat{\rho}_\mu = \hat{\sigma}_\mu^2/(\hat{\sigma}_\mu^2 + \hat{\sigma}_v^2)$.

Confidence intervals

[5] If the estimated σ_μ^2 is negative, which occurs when the true value is close to zero (Baltagi 2008, p. 20), it may be replaced by zero as suggested by Maddala and Mount (1973).

Confidence intervals at the desired significance level can be computed with the `estci` functions, and appropriately displayed with the `estcidisp` function. Both functions take as input an estimation output structure `estout` and the desired significance level, which defaults to 0.05 if not specified.

```
estcidisp(re);

Confidence Intervals at sig=0.05 (95%)
---------------------------------------------------------------------------
           lgsp |   Coefficient    Std. Error         Lower         Upper
---------------------------------------------------------------------------
          lpcap |      0.004439      0.023417     -0.041459      0.050336
            lpc |      0.310548      0.019805      0.271732      0.349365
           lemp |      0.729671      0.024920      0.680828      0.778513
          unemp |     -0.006172      0.000907     -0.007951     -0.004394
          CONST |      2.135411      0.133461      1.873831      2.396991
---------------------------------------------------------------------------
```

Robust standard errors

If we suspect that there exists heteroskedasticity in the residuals, we can compute a robust standard error estimation of the fixed and random effects models. Liang and Zeger (1986) and Arellano (1987) propose an extension of the White (1980) sandwich estimator for panel data models, whose asymptotic properties are studied by Hansen (2007) and Stock and Watson (2008). The correct standard errors should be computed as a clustered-robust standard errors using the observation groups as the different clusters.

$$\mathsf{VAR}(\hat{\beta}) = \frac{n}{n-1}\frac{N-1}{N-k}(\tilde{X}^\top\tilde{X})^{-1}\left[\sum_{i=1}^{n}\tilde{X}_i^\top e_i e_i^\top\tilde{X}_i\right](\tilde{X}^\top\tilde{X})^{-1}, \tag{19}$$

where, in the fixed effects estimation, $\tilde{X}$ is the within transformation of the explanatory variables, e are the residuals from the within regression, and the degrees of freedom correction $n/(n-1) \times N/(N-k)$ is usually applied. In a random effects estimation, $\tilde{X}$ is the quasi-time demeaning transformation of the explanatory variables, e the residuals from the random effects regression, and the degrees of freedom correction is $n/(n-1) \times (N-1)/(N-k)$.

The `panel` function allows robust standard errors estimation, both for fixed and random effects, by setting the option `vartype` to `robust`.

```
fer = panel(id, year, y, X, 'fe', 'vartype', 'robust');
fer.ynames = ynames;
fer.xnames = xnames;
estdisp(fer);

Panel: Fixed effects (within) (FE)

N = 816   n = 48   T = 17 (Balanced panel)
R-squared = 0.94134    Adj R-squared = 0.93742
```

```
Wald F(4, 47) = 395.612524 p-value = 0.0000
RSS = 1.111189 ESS = 90964.408970 TSS = 90964.408970
Standard errors robust to heteroskedasticity adjusted for 48 clusters

-----------------------------------------------------------------------
           lgsp |   Coefficient   Rob.Std.Err      t-stat    p-value
-----------------------------------------------------------------------
          lpcap |     -0.026150      0.061115     -0.4279    0.671
            lpc |      0.292007      0.062549      4.6684    0.000 ***
           lemp |      0.768159      0.082732      9.2849    0.000 ***
          unemp |     -0.005298      0.002528     -2.0952    0.042 **
-----------------------------------------------------------------------
```

Standard errors can be adjusted to a different cluster by setting the option `vartype` to `cluster`, and specifying the cluster variable to the option `clusterid`.[6]

3.2. Instrumental panels

The assumption of strict exogeneity of the independent variables, X, when they are uncorrelated with the disturbance, $\mathsf{E}(X_{it}, v_{it}) = 0$, implies that the basic panel data methods we have shown remain valid. However, there are many applications in which this assumption is untenable. In this case, when some of the regressors are endogenous, the fixed effects, between, and random effects estimators lose consistency and unbiasedness. Consequently, we can apply an instrumental variables (IV) two stage estimation to the fixed effects, between, and random effects models (Wooldridge 2010).

To apply this estimation method, we need a set of variables that are strictly exogenous, uncorrelated with the disturbance in all time periods, and relevant; i.e., correlated with the endogenous independent variables. These variables constitute the set of instrumental variables (IV).

For an application of instrumental panel data, we follow Baltagi and Levin (1992) and Baltagi, Griffin, and Xiong (2000) who estimate the demand for cigarettes using data from 46 U.S. states over the period 1963–1992.[7] We estimate the consumption, `log(c)`, measured as per capita sales, which depends on the price per pack, `log(price)`, per capita disposable income, `log(ndi)`, and the minimum price in neighbor states, `log(pimin)`.[8] We believe the `log(price)` is potentially endogenous, and use as instrumental variables the lags of the disposable income, `log(ndi_1)` and the lag of the minimum price `log(pimin_1)`.

```
load('CigarData')
y = log(c);
X = [log(price), log(ndi), log(pimin)];
Z = [log(ndi_1), log(pimin_1)];
ynames = {'lc'};
xnames = {'lprice', 'lndi', 'lpimin'};
znames = {'lndi_1', 'lpimin_1'};
```

[6] In fact, setting `vartype` to `robust` is equivalent to setting `vartype` to `cluster` and `clusterid` to `id`.

[7] The data is available in MATLAB format in the supplementary file `CigarData.mat`.

[8] The equation we estimate differs from the original one, which corresponds to a dynamic panel data model.

Instrumental panel models are estimated using the `ivpanel(id, time, y, X, Z, method, options)` function, where `Z` is the matrix of instruments – excluding the exogenous variables in `X` that are instruments of themselves and are automatically added by the function. A vector of indexes corresponding to the endogenous variables must be set in the `endog` option. `method` is a string that specifies the choice of instrumental panel data estimation method, among the following:

- `po`: For a pool estimation.
- `fe`: For a fixed effects (within) estimation.
- `be`: For a between effects estimation.
- `re`: For a random effects estimation.
- `ec`: For a error-components estimation (Baltagi 1981).

Two stage least squares

Instrumental panel data models are estimated by two stage least squares (2SLS). The first stage of the 2SLS estimation consists of estimating the independent variables, $\hat{X}$, by an OLS estimation of $\tilde{X}$ over $\tilde{H} = [\tilde{X}^{*}, \tilde{Z}]$, where $\tilde{X}^{*}$ are the exogenous variables in $\tilde{X}$, which are instruments of themselves, and $\tilde{Z}$ is the matrix of new instruments. For simplification, the tilde over the variables denotes the corresponding within, between or quasi-time demeaning transformation.

$$\hat{X} = \tilde{H}(\tilde{H}^{\top}\tilde{H})^{-1}\tilde{H}^{\top}\tilde{X}. \tag{20}$$

The second stage consists in estimating the coefficients, $\hat{\beta}$, using the predicted $\hat{X}$:

$$\hat{\beta}_{2SLS} = (\hat{X}^{\top}\tilde{X})^{-1}\hat{X}^{\top}\tilde{y}. \tag{21}$$

Wherever $\tilde{X}$ and $\tilde{H}$ correspond to the within, between, or quasi-time demeaning transformation of the variables, we are computing the corresponding fixed effects 2SLS (FE2SLS), between 2SLS (BE2SLS), and random effects 2SLS (RE2SLS).

Regarding statistical inference, the statistic of individual significance is normally distributed, while the statistic of joint significance is distributed as a χ^2 distribution with the corresponding degrees of freedom.

The `ivpanel` function implements the estimation of fixed, between and random effects two stage least squares instrumental panel data models in MATLAB:

```
ivfe = ivpanel(state, year, y, X, Z, 'fe', 'endog', 1);
ivfe.ynames = ynames;
ivfe.xnames = xnames;
ivfe.znames = znames;
estdisp(ivfe);
```

```
IV Panel: Fixed effects two stage least squares (FE2SLS)

N = 1334  n = 46  T = 29 (Balanced panel)
```

```
R-squared = 0.64064    Adj R-squared = 0.62722
Wald Chi2(3) = 1792.756633 p-value = 0.0000
RSS = 7.731114 ESS = 30699.227796 TSS = 30699.227796

----------------------------------------------------------------------
             lc |  Coefficient    Std. Error      z-stat    p-value
----------------------------------------------------------------------
         lprice |    -1.016355      0.249197     -4.0785    0.000 ***
           lndi |     0.537848      0.023033     23.3507    0.000 ***
         lpimin |     0.312372      0.228395      1.3677    0.171
----------------------------------------------------------------------
Endogenous: lprice
Instruments (exogenous): lndi lpimin
Instruments (new): lndi_1 lpimin_1
----------------------------------------------------------------------
```

Baltagi's error components estimator

Baltagi (1981) suggests an alternative error components two stage least squares (EC2SLS) estimation, based on a generalized two stage least squares estimator of the coefficients, $\hat{\beta}$, as for random effects using the following matrix of instruments:

$$A = [\tilde{H}, \bar{H}], \tag{22}$$

where $\tilde{H}$ corresponds to the within transformation of the instruments H, and $\bar{H}$ are the group means of the instruments. Then, the EC2SLS estimation is performed using A as the matrix of instruments in a random effects context.[9]

Consequently, EC2SLS incorporates more instruments than RE2SLS. Baltagi and Li (1992) show that both estimators are consistent and have the same limiting distributions, although it is worth noting that for small samples EC2SLS shows gains in efficiency. More recently, Baltagi and Liu (2009) present proofs to obtain the EC2SLS asymptotic properties with respect to RE2SLS.

The error components two stage least squares (EC2SLS) estimation can also be performed with the `ivpanel` by specifying the `'ec'` method:

```
ec2sls = ivpanel(state, year, y, X, Z, 'ec', 'endog', 1);
ec2sls.ynames = ynames;
ec2sls.xnames = xnames;
ec2sls.znames = znames;
estdisp(ec2sls);
```

```
Panel: Baltagi's error components two stage least squares (EC2SLS)

N = 1334  n = 46  T = 29 (Balanced panel)
```

[9] The instrument A is used when computing the 2SLS estimation, but the original H is used when estimating σ_v^2 and σ_1^2.

```
R-squared = 0.41686    Adj R-squared = 0.41554
Wald Chi2(3) = 1825.252894 p-value = 0.0000
RSS = 7.883472 ESS = 30699.075438 TSS = 30699.075438

----------------------------------------------------------------------
             lc |   Coefficient    Std. Error      z-stat    p-value
----------------------------------------------------------------------
         lprice |     -0.992679      0.235869     -4.2086    0.000 ***
           lndi |      0.536410      0.022356     23.9939    0.000 ***
         lpimin |      0.290388      0.215970      1.3446    0.179
          CONST |      2.995124      0.084198     35.5724    0.000 ***
----------------------------------------------------------------------
sigma_mu = 0.190101      rho_mu = 0.857278
 sigma_v = 0.077566     sigma_1 = 0.190646
   theta = 0.924449
----------------------------------------------------------------------
Endogenous: lprice
Instruments (exogenous): lndi lpimin
Instruments (new): lndi_1 lpimin_1
----------------------------------------------------------------------
```

3.3. Spatial panels

In recent years the econometrics literature has grown with topics related to the analysis of spatial relations using panel data models. The main reason is the availability of more complete data sets in which units characterized by spatial features are observed over time. In general, a spatial panel data set contains more information and less multicollinearity among the variables than a cross-section spatial counterpart – see Anselin (1988, 2010), Elhorst (2014b) and Arbia (2014) for an introduction to this literature.

In the context of cross-sectional models Kelejian and Prucha (1998) introduce a generalized spatial two-stage least squares estimator, Kelejian and Prucha (1999)[10] propose a generalized moments (GM) estimation method that is feasible for large n, while Anselin (1988) provides the ML (maximum likelihood) estimator. Drukker, Egger, and Prucha (2013) extend the model allowing for endogenous regressors. Most recently, Elhorst (2003, 2010) and Lee and Yu (2010) present the ML estimators of the spatial lag model as well as the error model extended to include fixed and random effects, solving the computational problems when the number of cross sectional units n is large. Kapoor *et al.* (2007), Mutl and Pfaffermayr (2011), and Piras (2013) generalize the GM procedure from cross-section to panel data and derive its properties.

In order to compute different estimators in spatial panel models, we consider the general

[10] Kelejian and Prucha (2004) extend the model to a system of equation spatially interrelated, while Kelejian and Prucha (2007, 2010) introduced a method robust to heteroscedasticity and autocorrelation in disturbances in a spatial autoregressive model.

spatial panel model:

$$\begin{aligned} y_{it} &= \lambda W y_{it} + \beta X_{it} + \beta_\lambda W X_{it} + \mu_i + \varepsilon_{it}, &(23)\\ \varepsilon_{it} &= \rho W \varepsilon_{it} + v_{it}. &(24)\end{aligned}$$

A spatial panel data model can include a spatial lag of the dependent variable, $W y_{it}$, a spatial lag in the error structure, $W\epsilon_{it}$, and a spatial lag in the explanatory variables, WX_{it}, whose coefficients are λ, ρ, and β_λ, respectively. Depending on the spatial lags they include the model receives a different name.

Procedures for estimating spatial panel data models in MATLAB are already available in LeSage and Pace (2009), using Bayesian methods, and in Elhorst (2014a), by maximum likelihood. In this toolbox, we implement the GM procedure for spatial panels, which allows the inclusion of additional endogenous covariates, and it is integrated with the rest of the toolbox, both regarding estimation and testing functions.[11]

In the case where only the spatial lag of the dependent variable is included, this spatial lag is endogenous and the estimation of the spatial model is performed as an instrumental variables estimation using the instruments suggested by Kelejian and Prucha (1998), $H = [X, WX, W^2X]$. If the model contains a spatial lag of the error structure, the estimation method is a GM estimation, and we refer the reader to Kapoor *et al.* (2007), Mutl and Pfaffermayr (2011), and Piras (2013) for a full explanation of the estimation methods and the corresponding moments conditions.

The application is based on the Munnell (1990) and Baltagi (2008) data of U.S. states production.[12]

```
load('MunnellData')
load('MunnellW')
y = log(gsp);
X = [log(pcap), log(pc), log(emp), unemp];
ynames = {'lgsp'};
xnames = {'lpcap', 'lpc', 'lemp', 'unemp'};
```

Spatial panel data models are estimated using the `spanel(id, time, y, X, W, method, options)`, where W is the $n \times n$ spatial weight matrix.[13] `method` can be one of the following:[14]

- `fe`: For a spatial fixed effects (within) estimation.
- `re`: For a spatial random effects estimation.

[11] These three packages work by taking the data as input and returning a structure with the results of the estimation as output. Although LeSage and Pace (2009) and Elhorst (2014a) use different functions for estimating models with different spatial lags, here all are condensed in a single `spanel` function which allows to estimate models by selecting which spatial lags to include. Despite this small difference, the user will find no difficulty in using the three packages if he wants to compare results using different estimation procedures.

[12] The Munnell (1990) data is available in MATLAB format in the supplementary file `MunnellData.mat`, while the W matrix comes from Millo and Piras (2012) and is available in the file `MunnellW.mat`.

[13] The function transforms the W matrix into a sparse matrix to take advantage of the computational speed improvements of MATLAB when working with sparse matrices.

[14] As for now, spatial panels are only available for balanced panels, since the methods for unbalanced ones are still in their early stages.

- `ec`: For the Baltagi and Liu (2011) spatial error components estimation of the model with a spatial lag of the dependent varaible.

The different spatial lags can be included by setting the following options:

- `slagy`: If set to 1 includes a spatial lag of the dependent variables.
- `slagerror`: If set to 1 includes a spatial lag of the error structure.
- `slagX`: A vector of indexes specifying the explanatory variables for which a spatial lag should be added.

Estimating a model with a spatial lag in the dependent variable and a spatial lag in the error structure, usually denoted as SARAR (spatial autoregressive with additional autoregressive error structure), is straightforwardly performed with the `spanel` function:

```
sarar = spanel(id, year, y, X, W, 're', 'slagy', 1, 'slagerror', 1);
sarar.ynames = ynames;
sarar.xnames = xnames;
estdisp(sarar);
```

```
Spatial Panel: Random effects spatial two stage least squares (RES2SLS)

N = 816  n = 48  T = 17 (Balanced panel)
R-squared = 0.99123
Wald Chi2(5) = 15681.075028 p-value = 0.0000
RSS = 7.461059

----------------------------------------------------------------------
           lgsp |   Coefficient    Std. Error      z-stat    p-value
----------------------------------------------------------------------
          lpcap |      0.046326      0.022686      2.0420    0.041 **
            lpc |      0.267972      0.020473     13.0891    0.000 ***
           lemp |      0.720149      0.024939     28.8769    0.000 ***
          unemp |     -0.005233      0.000978     -5.3497    0.000 ***
         W*lgsp |      0.022307      0.013542      1.6472    0.100 *
          CONST |      2.006880      0.168351     11.9208    0.000 ***
----------------------------------------------------------------------
            rho |      0.325480      0.001131    287.8803    0.000 ***
----------------------------------------------------------------------
 sigma_v = 0.033625     sigma_1 = 0.305323
   theta = 0.889872
----------------------------------------------------------------------
Endogenous: W*lgsp
----------------------------------------------------------------------
```

The `spanel` function also allows to perform spatial panel estimation when one of the explanatory variables is endogenous. This is performed by including a vector of indexes of the endogenous variables in the option `endog`, and passing the matrix of new instruments to the option `inst`. For example, if we assume that the public capital `log(pcap)` is exogenous and we want to instrument it using the highway and the water components of the public capital, `log(hwy)` and `log(water)`:

```
Z = [log(hwy), log(water)];
sarfe = spanel(id, year, y, X, W, 'fe', 'slagy', 1, 'slagerror', 1,...
    'endog', 1, 'inst', Z);
sarfe.ynames = ynames;
sarfe.xnames = xnames;
estdisp(sarfe);
```

```
Spatial Panel: Fixed effects spatial two stage least squares (FES2SLS)

N = 816  n = 48  T = 17 (Balanced panel)
R-squared = 0.98248
Wald Chi2(5) = 7450.217570 p-value = 0.0000
RSS = 3292.934466

----------------------------------------------------------------------
           lgsp |   Coefficient    Std. Error      z-stat    p-value
----------------------------------------------------------------------
          lpcap |      0.026432      0.035201      0.7509    0.453
            lpc |      0.188595      0.025652      7.3521    0.000 ***
           lemp |      0.713135      0.031572     22.5875    0.000 ***
          unemp |     -0.004263      0.001074     -3.9705    0.000 ***
         W*lgsp |      0.124480      0.024919      4.9954    0.000 ***
----------------------------------------------------------------------
            rho |      0.338480      0.001132    299.0594    0.000 ***
----------------------------------------------------------------------
Endogenous: lpcap W*lgsp
----------------------------------------------------------------------
```

4. Tests

In this section we describe the implementation of several canonical tests for the panel data regression models presented previously. Specification tests in panel data involves testing for poolability, individual effects and the Hausman test to select the efficient estimator between fixed and random effects models. In addition, we provide a suite of serial correlation and cross-sectional dependence tests. Finally, we consider as the usual diagnostic checks an overidentification test for validity of instruments in instrumental panels and tests for spatial autocorrelation in spatial panels. Appropriate corrections for heteroskedasticity and unbalanced panels for these tests are applied when available.

All test functions require as input an estimation output structure, `estout`, from a panel estimation and return a `testout` structure, described in Section 2, that can be displayed in a suitable way using the `testdisp` function.

4.1. Testing linear hypotheses

Linear hypotheses of the form $H_0 : R\beta = r$ can be tested with the standard Wald joint significance test, using the `waldsigtest` function and specifying the `R` and `r` matrices of the null hypothesis to be tested.

```
R = [1 0 0 0 0; 0 1 0 0 0];
r = [0; 0];
wald = waldsigtest(re, R, r);
testdisp(wald);

Wald joint significance test

Chi2(2) = 250.337223
p-value = 0.0000
```

4.2. Testing poolability

`pooltest` tests the hypothesis that the population parameters are the same across individuals. Therefore we want to test the stability of the coefficients, $H_0 : \beta_i = \beta$ for all i, in Equation 1. It is a standard F test based on a comparison between the model estimated for the complete sample and a model that estimates an equation for each individual (Baltagi 2008).

```
pool = pooltest(re);
testdisp(pool);

Test of poolability

H0: Stability of coefficients
F(282,528) = 33.829171
p-value = 0.0000
```

4.3. Testing individual effects

The test for individual effects contrasts the existence of different time invariant specific effects based on the results of the pooling model. `effectsftest` performs the Chow F test for individual effects as in Baltagi (2008). Under the null hypothesis that there are no individual effects, $\mu_i = 0 \ \forall i$, the restricted model comes from an OLS pooling estimation, while the unrestricted model follows the fixed effects estimation.

```
effF = effectsftest(fe);
testdisp(effF)

F test of individual effects

H0: All mu_i = 0
F(47,764) = 75.820406
p-value = 0.0000
```

`bpretest` implements the Baltagi and Li (1990) version of the Lagrange multiplier (LM) test of individual effects proposed by Breusch and Pagan (1980). This test contrasts the existence of individual effects by checking its variance that under the null hypothesis of no individual effects is equal to zero, and the LM statistic is distributed as a χ^2_1.

```
bpre = bpretest(re);
testdisp(bpre);
```

```
Breusch-Pagan's LM test for random effects

Baltagi and Li (1990) version of the Breusch and Pagan (1980) test
H0: sigma2_mu = 0
  LM = 4134.960740 ~ Chi2(1)
p-value = 0.0000
```

4.4. Testing fixed vs. random effects

In order to determine the correct specification of the model, fixed versus random effects, it is necessary to check the correlation between the individual effects and the regressors. When the individual effects and the explanatory variables are correlated: $\mathsf{COV}(X_{it}, \mu_i) \neq 0$, the fixed effects model provides an unbiased estimator, otherwise a feasible GLS estimator in a random effects model is an efficient estimator.

`hausmantest` computes the Hausman test (Hausman 1978) that compares the GLS estimator of the random effects model, $\hat{\beta}_{re}$, and the within estimator in the fixed effects model, $\hat{\beta}_{fe}$, both of which are consistent under the null hypothesis. Under the alternative, only the GLS estimator of random effects is consistent. Therefore, the statistics is based on the difference between both estimators $H_0 : \beta_{fe} - \beta_{re} = 0$, and it is computed as:

$$H = (\hat{\beta}_{fe} - \hat{\beta}_{re})^\top \mathsf{VAR}(\hat{\beta}_{fe} - \hat{\beta}_{re})^{-1}(\hat{\beta}_{fe} - \hat{\beta}_{re}),$$

where, under the assumption of homoskedasticity:

$$\mathsf{VAR}(\hat{\beta}_{fe} - \hat{\beta}_{re}) = \mathsf{VAR}(\hat{\beta}_{fe}) - \mathsf{VAR}(\hat{\beta}_{re}).$$

For n fixed and T large, both estimators tend to similar values, with their difference converging to zero, and Hausman's test is unnecessary. However, in applications where n is relatively large with respect to T, it can be used to choose between estimators.

The input of the `hausmantest` function requires the output structures of the two estimations to be compared.

```
hausman = hausmantest(fe, re);
testdisp(hausman);
```

```
Hausman's test of specification

----------------------------------------------------------------------
```

```
        Varname |          A:FE          B:RE    Coef. Diff    S.E. Diff
------------------------------------------------------------------------
          lpcap |     -0.026150      0.004439     -0.030588     0.017109
            lpc |      0.292007      0.310548     -0.018542     0.015452
           lemp |      0.768159      0.729671      0.038489     0.016867
          unemp |     -0.005298     -0.006172      0.000875     0.000393
------------------------------------------------------------------------
A is consistent under H0 and H1 (A = FE)
B is consistent under H0        (B = RE)
H0: coef(A) - coef(B)  = 0
H1: coef(A) - coef(B) != 0
      H = 9.525416 ~ Chi2(4)
p-value = 0.0492
```

In case of a spatial panel data model with a spatial lag in the error structure, the spatial Hausman test described in Mutl and Pfaffermayr (2011) is performed by passing the spatial estimation output structures to the `hausmantest` function.

The Mundlak (1978) approach suggests estimating the following regression by GLS:

$$y_{it} = \alpha + X_{it}\beta + \bar{X}_i\gamma + \mu_i + v_{it}, \qquad i = 1, \dots, n, \quad t = 1, \dots, T_i, \tag{25}$$

where $\bar{X}_i$ are the group means of the variables. Then, a test can be performed by computing a Wald joint significance test on γ, under the null hypothesis of random effects, $H_0 : \gamma = 0$. This approach is computationally more stable in finite samples and can be estimated with robust standard errors (Wooldridge 2010).

```
mundlak = mundlakvatest(fe);
testdisp(mundlak);
```

```
Mundlak's variable addition test for fixed or random effects

H0: Group means are zero. Random effects.
Chi2(4) = 9.718105
 p-value = 0.0455
```

4.5. Testing serial correlation

In linear panel data models it is necessary to identify serial correlation in the error term because it biases the standard errors and causes loss of efficiency. We present tests for serial correlation in random and fixed effects models.

`woolserialtest` performs the Wooldridge's test (Wooldridge 2010) for the null hypothesis of no serial correlation in the error term of a fixed effects model. Under the null hypothesis of no serial correlation in the errors, v_{it}, the time demeaned errors of a within regression are negatively serially correlated, with correlation $\rho = -1/(T-1)$. Thus, a test of serial correlation can be performed by regressing the within estimation residuals, $\hat{v}_{it}$, over their lag, $\hat{v}_{i,t-1}$:

$$\hat{v}_{it} = \alpha + \rho\hat{v}_{it} + \epsilon_{it},$$

and testing whether $\hat{\rho} = -1/(T-1)$, using a Wald test with clustered standard errors.

```
woolfe = woolserialtest(fe);
testdisp(woolfe);
```

```
Wooldridge's test for serial correlation

H0: Corr(res_{T-1}, res_T) = rho. No serial correlation
rho = -1/(T-1) = -0.062500
F(1,47) = 680.299012
p-value = 0.0000
```

In the context of a random effects model `blserialtest` performs the Lagrange multiplier test for first-order serially correlated errors and random effects proposed by Baltagi and Li (1990), as an extension to Breusch and Pagan (1980). This test contrasts the joint null hypothesis of serial correlated and random individual effects. The LM test is based on the OLS residuals and it is asymptotically distributed as a χ^2_2.

```
blre = blserialtest(re);
testdisp(blre);
```

```
Baltagi and Li's test for serial correlation and random effects

H0: No random effects and no serial correlation.
H1: Random effects or serial correlation.
Chi2(2) = 4187.596596
p-value = 0.0000
```

4.6. Testing cross-sectional dependence

Cross-sectional dependence in the errors may arise because of the presence of common shocks or when the estimated models present spatial dependence in the disturbances. Cross-sectional dependence results in the inefficiency of the usual estimators and an invalid inference when using the standard covariance matrix. This indicates that testing for cross-sectional dependence is important in fitting panel data models.

`pesarancsdtest` implements the Pesaran (2004) cross-sectional dependence (CD) test for balanced and unbalanced panels. Under the null hypothesis of no cross-sectional dependence, the Pesaran's CD statistic is asymptotically distributed as a standard normal.

```
pesaran = pesarancsdtest(fe);
testdisp(pesaran);
```

```
Pesaran's test of cross sectional dependence

H0: Corr(res_{it}, res_{jt}) = 0 for i != j
     CD = 30.368501
p-value = 0.0000
```

4.7. Testing overidentification

To evaluate the validity of the instruments in instrumental panels we perform an overidentification test. The function `sarganoitest` performs the Sargan (1958) test of overidentification restrictions regressing the residuals of the instrumental estimation on all the instruments, including the exogenous variables that are instruments of themselves. Under the null hypothesis that instruments are uncorrelated with the error term, validity of the overidentifying restrictions, the statistic is distributed as a χ^2_r where r is the number of overidentifying restrictions. The input of the `sarganoitest` function must be an estimation output structure from an instrumental panel.

```
sargan = sarganoitest(ivfe);
testdisp(sargan);
```

```
Sargan's test of overidentification

H0: Instruments are uncorrelated with the error term
  Score = 25.520199 ~ Chi2(1)
p-value = 0.0000
```

4.8. Testing spatial autocorrelation

The function `bsjksatest` implements the join Lagrange multiplier test for testing serial correlation, spatial autocorrelation and random effects in spatial panels by Baltagi, Song, Jung, and Koh (2007). The test is based on the OLS residuals and the `W` matrix and under the null hypothesis of no spatial autocorrelation, no serial error correlation and no random effects, it is distributed as a χ^2_3. The input of the `bsjksatest` function must be an estimation output structure from a spatial panel.

```
bsjk = bsjksatest(sarar);
testdisp(bsjk);
```

```
Baltagi, Song, Jung and Koh's test for serial correlation,
  spatial autocorrelation and random effects

H0: No spatial autocorrelation, no serial error correlation and no re.
H1: Spatial autocorrelation or serial error correaltion or random effects.
Chi2(3) = 4290.422435
p-value = 0.0000
```

5. Numerical checks

Numerical checks against other commercial and free software are performed by comparing the panel data estimation results from this **Panel Data Toolbox** in MATLAB (The MathWorks Inc. 2015) and results reported by Stata (StataCorp 2015) and R (R Core Team 2016).[15]

[15] The code of this section for MATLAB, Stata and R are available in the files `NC_MATLAB.m`, `NC_Stata.do` and `NC_R.R` respectively.

		Coefficient			*t* statistic		
		MATLAB	Stata	R	MATLAB	Stata	R
Fixed	`lpcap`	−0.026150	−0.026150	−0.026150	−0.9017	−0.9017	−0.9017
	`lpc`	0.292007	0.292007	0.292007	11.6246	11.6246	11.6246
	`lemp`	0.768159	0.768159	0.768159	25.5273	25.5273	25.5273
	`unemp`	−0.005298	−0.005298	−0.005298	−5.3582	−5.3582	−5.3582
Between	`lpcap`	0.179365	0.179365	0.179365	2.4922	2.4922	2.4922
	`lpc`	0.301954	0.301954	0.301954	7.2201	7.2201	7.2201
	`lemp`	0.576127	0.576127	0.576127	10.2196	10.2196	10.2196
	`unemp`	−0.003890	−0.003890	−0.003890	−0.3926	−0.3926	−0.3926
	`CONST`	1.589444	1.589444	1.589444	6.8222	6.8222	6.8222
Random	`lpcap`	0.004439	0.004439	0.004439	0.1895	0.1896	0.1895
	`lpc`	0.310548	0.310548	0.310548	15.6805	15.6805	15.6805
	`lemp`	0.729671	0.729671	0.729671	29.2803	29.2803	29.2803
	`unemp`	−0.006172	−0.006172	−0.006172	−6.8033	−6.8033	−6.8033
	`CONST`	2.135411	2.135411	2.135411	16.0002	16.0002	16.0002

Table 1: Comparison of estimated coefficients and t statistics for panel data against Stata and R.

		Coefficient			*t* statistic		
		MATLAB	Stata	R	MATLAB	Stata	R
Fixed	`lprice`	−1.016355	−1.016359	−1.016355	−4.0785	−4.0785	−4.0785
	`lndi`	0.537848	0.537848	0.537848	23.3507	23.3508	23.3507
	`lpimin`	0.312372	0.312376	0.312372	1.3677	1.3677	1.3677
Random	`lprice`	−1.007113	−1.007117	−1.007113	−4.0715	−4.0716	−4.0715
	`lndi`	0.537473	0.537474	0.537473	23.3398	23.3398	23.3398
	`lpimin`	0.303567	0.303571	0.303567	1.3407	1.3407	1.3407
	`CONST`	2.992121	2.992121	2.992121	34.9268	34.9268	34.9268
Error components	`lprice`	−0.992679	−0.992681	−0.992679	−4.2086	−4.2086	−4.2086
	`lndi`	0.536410	0.536411	0.536410	23.9939	23.9939	23.9939
	`lpimin`	0.290388	0.290389	0.290388	1.3446	1.3446	1.3446
	`CONST`	2.995124	2.995124	2.995124	35.5724	35.5724	35.5724

Table 2: Comparison of estimated coefficients and t statistics for instrumental panel data against Stata and R.

Results for the basic panel data models – fixed, between and random – estimations using the MATLAB `panel` function, and the results reported by Stata `xtreg` function, and the `plm` function from the R package **plm** by Croissant and Millo (2008), are reported in Table 1. Results show that there are no differences in the estimated coefficients and t statistics between the three programs.

Numerical checks for the instrumental variables panel data models of fixed effects, random effects, and Baltagi's error components using the MATLAB `ivpanel` function, the Stata `xtivreg` function, and `plm` function from the R package **plm** are reported in Table 2. Again, results are equal regardless of the software, although there is a slightly difference in the last decimal between Stata and the other two.

Spatial panel estimations using the MATAB function `spanel` are checked against the R package **splm** by Millo and Piras (2012), using the `spgm` function, which performs a GM implementation. Since a large variety of models can be computed for spatial panels depending on the

		Coefficient		*t* statistic	
		MATLAB	R	MATLAB	R
Fixed	lpcap	−0.020583	−0.020583	−0.7660	−0.7660
	lpc	0.193687	0.193687	7.5842	7.5842
	lemp	0.729175	0.729175	24.0058	24.0058
	unemp	−0.003700	−0.003700	−3.6154	−3.6154
	W*lgsp	0.132709	0.132709	5.3963	5.3963
	rho	0.325480	0.325480	9.6798	9.6798
Random	lpcap	0.046326	0.046326	2.0420	2.0420
	lpc	0.267972	0.267972	13.0891	13.0891
	lemp	0.720149	0.720149	28.8769	28.8769
	unemp	−0.005233	−0.005233	−5.3497	−5.3497
	W*lgsp	0.022307	0.022307	1.6472	1.6472
	CONST	2.006880	2.006880	11.9208	11.9208
	rho	0.325480	0.325480	9.6798	9.6798

Table 3: Comparison of estimated coefficients and *t* statistics for spatial panel data against R.

spatial lags we assume, we perform the numerical checks of a spatial SARAR model, which includes a spatial lag of the dependent variable and a spatial lag of the error structure, both with fixed and random effects. Although different interpretations of the literature as well as on the choice of techniques when implementing spatial econometrics lead to some differences in the results (Bivand and Piras 2015), results in Table 3 reveal no differences in the estimated coefficients and *t* statistics between MATLAB and R.

6. Conclusions

The new **Panel Data Toolbox** covers a wide variety of balanced and unbalanced panel data models in an organized environment for MATLAB. Estimation methods include fixed, between and random effects, as well as instrumental and spatial panels, and the full set of relevant tests for testing poolability, individual effects, serial correlation, cross-sectional dependence, overidentification and spatial autocorrelation.

Numerical checks show the consistency of the results, as the estimated coefficients and *t* statistics are equal to those reported by Stata and R for panel, instrumental panels and spatial panel data methods. This positions the new toolbox as a valid self-contained package for panel data econometrics in MATLAB.

Since the code is freely available in an open source repository on GitHub at https://github.com/javierbarbero/PanelDataMATLAB, under the GNU General Public License version 3, users will benefit from the review, collaboration and contributions from the community, and can check the syntax to learn how the theoretical formulas of econometrics can be translated into code.

Acknowledgments

We thank the editor and two anonymous referees for helpful comments and suggestions, as well as users of the previous release of the toolbox. This research was supported by the Spanish Ministry of Science and Innovation under research grant (ECO2010-21643). Javier Barbero acknowledges financial support from the Spanish Ministry of Education (AP2010-1401).

References

Anselin L (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publisher, Dordrecht. doi:10.1007/978-94-015-7799-1.

Anselin L (2010). "Thirty Years of Spatial Econometrics." *Papers in Regional Science*, **89**(1), 3–25. doi:10.1111/j.1435-5957.2010.00279.x.

Arbia G (2014). *A Primer for Spatial Econometrics: With Applications in R*. Palgrave Macmillan. doi:10.1057/9781137317940.

Arellano M (1987). "Practitioners' Corner: Computing Robust Standard Errors for Within-Groups Estimators." *Oxford Bulletin of Economics and Statistics*, **49**(4), 431–434. doi:10.1111/j.1468-0084.1987.mp49004006.x.

Baltagi BH (1981). "Simultaneous Equations with Error Components." *Journal of Econometrics*, **17**(2), 189–200. doi:10.1016/0304-4076(81)90026-9.

Baltagi BH (2008). *Econometric Analysis of Panel Data*. 4th edition. John Wiley & Sons, United Kingdom.

Baltagi BH, Griffin JM, Xiong W (2000). "To Pool or Not to Pool: Homogeneous Versus Hetergeneous Estimations Applied to Cigarette Demand." *The Review of Economics and Statistics*, **82**(1), 117–126. doi:10.1162/003465300558551.

Baltagi BH, Levin D (1992). "Cigarette Taxation: Raising Revenues and Teducing Consumption." *Structural Change and Economic Dynamics*, **3**(2), 321–335. doi:10.1016/0954-349x(92)90010-4.

Baltagi BH, Li Q (1990). "A Lagrange Multiplier Test for the Error Components Model with Incomplete Panels." *Econometric Reviews*, **9**(1), 103–107. doi:10.1080/07474939008800180.

Baltagi BH, Li Q (1992). "A Note on the Estimation of Simultaneous Equations with Error Components." *Econometric Theory*, **8**(1), 113–119. doi:10.1017/s0266466600010768.

Baltagi BH, Liu L (2009). "A Note on the Application of EC2SLS and EC3SLS Estimators in Panel Data Models." *Statistics & Probability Letters*, **79**(20), 2189–2192. doi:10.1016/j.spl.2009.07.014.

Baltagi BH, Liu L (2011). "Instrumental Variable Estimation of a Spatial Autoregressive Panel Model with Random Effects." *Economics Letters*, **111**(2), 135–137. doi:10.1016/j.econlet.2011.01.016.

Baltagi BH, Song SH, Jung BC, Koh W (2007). "Testing for Serial Correlation, Spatial Autocorrelation and Random Effects Using Panel Data." *Journal of Econometrics*, **140**(1), 5–51. doi:10.1016/j.jeconom.2006.09.001.

Bivand R, Piras G (2015). "Comparing Implementations of Estimation Methods for Spatial Econometrics." *Journal of Statistical Software*, **63**(18), 1–36. doi:10.18637/jss.v063.i18.

Breusch TS, Pagan AR (1980). "The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics." *The Review of Economic Studies*, **47**(1), 239–253. doi:10.2307/2297111.

Croissant Y, Millo G (2008). "Panel Data Econometrics in R: The **plm** Package." *Journal of Statistical Software*, **27**(2), 1–43. doi:10.18637/jss.v027.i02.

Drukker DM, Egger P, Prucha IR (2013). "On Two-Step Estimation of a Spatial Autoregressive Model with Autoregressive Disturbances and Endogenous Regressors." *Econometric Reviews*, **32**(5–6), 686–733. doi:10.1080/07474938.2013.741020.

Elhorst JP (2003). "Unconditional Maximum Likelihood Estimation of Dynamic Models for Spatial Panels." *Research Report 03C27*, University of Groningen, Research Institute SOM (Systems, Organisations and Management).

Elhorst JP (2010). "Applied Spatial Econometrics: Raising the Bar." *Spatial Economic Analysis*, **5**(1), 9–28. doi:10.1080/17421770903541772.

Elhorst JP (2014a). "MATLAB Software for Spatial Panels." *International Regional Science Review*, **37**(3), 389–405. doi:10.1177/0160017612452429.

Elhorst JP (2014b). *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels.* Springer-Verlag.

Fuller WA, Battese GE (1973). "Transformations for Estimation of Linear Models with Nested-Error Structure." *Journal of the American Statistical Association*, **68**(343), 626–632. doi:10.1080/01621459.1973.10481396.

Fuller WA, Battese GE (1974). "Estimation of Linear Models with Crossed-Error Structure." *Journal of Econometrics*, **2**(1), 67–78. doi:10.1016/0304-4076(74)90030-x.

Hansen CB (2007). "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When T Is Large." *Journal of Econometrics*, **141**(2), 597–620. doi:10.1016/j.jeconom.2006.10.009.

Hausman JA (1978). "Specification Tests in Econometrics." *Econometrica*, **46**(6), 1251–1271. doi:10.2307/1913827.

Kapoor M, Kelejian HH, Prucha IR (2007). "Panel Data Models with Spatially Correlated Error Components." *Journal of Econometrics*, **140**(1), 97–130. doi:10.1016/j.jeconom.2006.09.004.

Kelejian HH, Prucha IR (1998). "A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances." *The Journal of Real Estate Finance and Economics*, **17**(1), 99–121. doi:10.1023/A:1007707430416.

Kelejian HH, Prucha IR (1999). "A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model." *International Economic Review*, **40**(2), 509–533. doi:10.1111/1468-2354.00027.

Kelejian HH, Prucha IR (2004). "Estimation of Simultaneous Systems of Spatially Interrelated Cross Sectional Equations." *Journal of Econometrics*, **118**(1–2), 27–50. doi:10.1016/s0304-4076(03)00133-7.

Kelejian HH, Prucha IR (2007). "HAC Estimation in a Spatial Framework." *Journal of Econometrics*, **140**(1), 131–154. doi:10.1016/j.jeconom.2006.09.005.

Kelejian HH, Prucha IR (2010). "Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances." *Journal of Econometrics*, **157**(1), 53–67. doi:10.1016/j.jeconom.2009.10.025.

Lee L, Yu J (2010). "Estimation of Spatial Autoregressive Panel Data Models with Fixed Effects." *Journal of Econometrics*, **154**(2), 165–185. doi:10.1016/j.jeconom.2009.08.001.

LeSage J, Pace RK (2009). *Introduction to Spatial Econometrics*. Chapman and Hall/CRC.

Liang KY, Zeger SL (1986). "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, **73**(1), 13–22. doi:10.2307/2336267.

Maddala GS, Mount TD (1973). "A Comparative Study of Alternative Estimators for Variance Components Models Used in Econometric Applications." *Journal of the American Statistical Association*, **68**(342), 324–328. doi:10.1080/01621459.1973.10482427.

Millo G, Piras G (2012). "**splm**: Spatial Panel Data Models in R." *Journal of Statistical Software*, **47**(1), 1–38. doi:10.18637/jss.v047.i01.

Mundlak Y (1978). "On the Pooling of Time Series and Cross Section Data." *Econometrica*, **46**(1), 69–85. doi:10.2307/1913646.

Munnell AH (1990). "Why Has Productivity Growth Declined? Productivity and Public Investment." *New England Economic Review*, **January/February**, 3–22.

Mutl J, Pfaffermayr M (2011). "The Hausman Test in a Cliff and Ord Panel Model." *The Econometrics Journal*, **14**(1), 48–76. doi:10.1111/j.1368-423x.2010.00325.x.

Pesaran MH (2004). "General Diagnostic Tests for Cross Section Dependence in Panels." *Cambridge Working Papers in Economics 0435*, Faculty of Economics, University of Cambridge.

Piras G (2013). "Efficient GMM Estimation of a Cliff and Ord Panel Data Model with Random Effects." *Spatial Economic Analysis*, **8**(3), 370–388. doi:10.1080/17421772.2013.804628.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sargan JD (1958). "The Estimation of Economic Relationships Using Instrumental Variables." *Econometrica*, **26**(3), 393–415. doi:10.2307/1907619.

StataCorp (2015). *Stata Statistical Software: Release 14*. StataCorp LP, College Station. URL http://www.stata.com/.

Stock JH, Watson MW (2008). "Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression." *Econometrica*, **76**(1), 155–174. doi:10.1111/j.0012-9682.2008.00821.x.

Swamy PAVB, Arora SS (1972). "The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models." *Econometrica*, **40**(2), 261–275. doi:10.2307/1909405.

The MathWorks Inc (2015). *MATLAB – The Language of Technical Computing, Version R2015a (8.5).* Natick, Massachusetts. URL http://www.mathworks.com/products/matlab/.

Wallace TD, Hussain A (1969). "The Use of Error Components Models in Combining Cross Section with Time Series Data." *Econometrica*, **37**(1), 55–72. doi:10.2307/1909205.

White H (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, **48**(4), 817–838. doi:10.2307/1912934.

Wooldridge JM (2010). *Econometric Analysis of Cross Section and Panel Data.* 2nd edition. The MIT Press.

Affiliation:

Inmaculada C. Álvarez, Javier Barbero, José L. Zofío
Department of Economics
Universidad Autónoma de Madrid
28049 Madrid, Spain
E-mail: inmaculada.alvarez@uam.es, javier.barbero@uam.es, jose.zofio@uam.es
URL: http://www.paneldatatoolbox.com/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
January 2017, Volume 76, Issue 6 *Submitted:* 2013-10-29
doi:10.18637/jss.v076.i06 *Accepted:* 2015-11-07