

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 $i = 1, 2, \ldots, n$, 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:

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

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.

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/516225235014010112