A Panel Data Toolbox for MATLAB

Inmaculada C. Álvarez, Javier Barbero and José L. Zofío

Working Paper 05/2013
A Panel Data Toolbox for MATLAB

Inmaculada C. Álvarez  Javier Barbero  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 panel data analysis. The package includes code for the standard fixed, between and random effects estimation methods, as well as for the existing instrumental panel and new spatial panel. 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 in order to show the validity of the results.

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

JEL codes: C21, C23, C26.

1. Introduction

Panel data econometrics have grown in importance over the past decades due to increase availability of data related to units that are observed over several periods of time. Panel data econometric methods are available in Stata and R, but there is a lack of a full set of functions for MATLAB, by The MathWorks, Inc. (2013).

The Panel Data Toolbox introduces such set of functions, including estimation methods for the standard fixed, between and random effects models, as well as instrumental panel data models, including the error components by Baltagi (1981) and Baltagi and Liu (2009), and, finally, existing and new spatial panel data, 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.

Spatial econometrics in MATLAB can be estimated using the LeSage and Pace (2009) Econometrics Toolbox, which uses maximum likelihood and bayesian methods, and Elhorst (2011) using maximum likelihood methods. In the new Panel Data Toolbox we use a two stage instrumental variables method to estimate spatial panels with fixed, between and random effects, as well as the error components model, following Baltagi and Liu (2011).

Panel Data Toolbox is available as free software and can be downloaded from http://www.paneldatatoolbox.com, with all the supplementary material (data and source code) to replicate all the results presented in this paper.

The paper is organized as follows. Section 3 presents the Panel data models with fixed, between and random effects. Instrumental panel data models are illustrated in Section 4. Spatial panels are covered in Section 5. Numerical checks against Stata and R are presented in Section 6. Finally, Section 7 concludes.
2. Data and structures

Panel data contains units (individuals, firms, countries, regions, etc.) that are observed over several periods of time. Units are usually denoted by \( i = 1, 2, \ldots, n \), and time periods by \( t = 1, 2, \ldots, T \). In this paper we deal only with the case of balanced panel data, those in which all units are observed over the same periods of time. Then, the total number of observations in the panel is \( N = nT \).

Data are managed as regular MATLAB vectors and matrices, constituting the inputs of the estimation functions. Observations are expected to be ordered first by units and then by time period. All estimation functions return a structure \texttt{estoutput} that contains properties with the estimation results as well as the input used to generate that output. Properties 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., \texttt{estprint}) or plot graphs (e.g., \texttt{estplot}).

Some of the properties of the \texttt{estoutput} structure are the following:\footnote{For a full list see the help of the function typing \texttt{help estoutput} in MATLAB.}

- \( y \) 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 (including the constant term).
- \( \texttt{coef} \), \( \texttt{varcoef} \) and \( \texttt{stderr} \): estimated coefficients, estimated covariance matrix, and estimated standard errors.
- \( \texttt{yhat} \) and \( \texttt{res} \): fitted values and residuals.
- \( \texttt{statistic} \), \( \texttt{df_statistic} \) and \( \texttt{p_statistic} \): statistic of individual significance, degrees of freedom of the statistic, and the corresponding \( p \) value.

3. Panel data models

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

\[
y_{it} = \alpha + X_{it}\beta + \mu_i + v_{it} \quad \forall i = 1, \ldots, n, \quad t = 1, \ldots, T,
\]

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

As a classic application we use 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, \( \log(\text{gsp}) \), and the explanatory ones are public capital, \( \log(\text{pcap}) \), private capital, \( \log(\text{pc}) \), employment, \( \log(\text{emp}) \), and the unemployment rate, \( \log(\text{unemp}) \).\footnote{Munnell (1990) data are available in MATLAB format in the supplementary file \texttt{MunnellData.mat}.}
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 \( \text{dvarnames} \) and \( \text{ivarnames} \) are cell arrays of strings that contain the name of the variables that are subsequently used when printing the results of the estimation.

Panel data models are estimated using the \texttt{panel(y, X, T, method, options)} function, where \( y \) is the vector of the dependent variable, \( X \) is the matrix of explanatory variables, \( T \) is the number of time periods per entity, and \( \text{method} \) is a string that specifies the panel data estimation method to be used among the following:

- \( \hat{\text{po}} \): for a pool estimation.
- \( \hat{\text{fe}} \): for a fixed effects (within) estimation.
- \( \hat{\text{be}} \): for a between effects estimation.
- \( \hat{\text{re}} \): for a random effects estimation

These estimation methods are explained in the following sections. \( \text{options} \) is an optional parameter to specify alternative estimation choices.

3.1. Fixed effects model

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

\[
\hat{\beta}_{fe} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{y},
\]

(2)

where \( \tilde{y} = y - \bar{y} \) and \( \tilde{X} = X - \bar{X} \) are the transformed variables in deviations from the group mean. It is called “within” estimator because it takes into account the variations in each group. This estimator is unbiased and consistent when both \( n \) and \( T \) are large. Statistical inference is generally based on the asymptotic variance covariance matrix:

\[
\text{VAR}(\hat{\beta}_{fe}) = S^2(\tilde{X}^T \tilde{X})^{-1},
\]

(3)

where \( S^2 \) denotes the residual variance: \( S^2 = (e^T e)/(n(T - 1) - k + 1) \), with residuals \( e = y - (X \hat{\beta}_{fe} + \alpha + \mu) \).
Finally, inference can be performed using the standard tests. The individual significance statistic is distributed as a $t$-student with $n(T - 1) - k + 1$ degrees of freedom under homoscedasticity, while the $F$ statistic of joint significance is:

$$F = \frac{Wald}{k - 1} \sim F_{k-1, n(T-1)-k+1}$$  \hspace{1cm} (4)$$

The goodness of fit is measured with the R-squared: $R^2 = 1 - \frac{(e^\top e)}{(\tilde{y}\tilde{y})}$, and the adjusted R-squared $\bar{R}^2 = 1 - \frac{(N - 1)/(N - k - n)(1 - R^2)}{N}$. The test for individual effects is the Chow test proposed in Baltagi (2008):

$$F = \frac{(RRSS - URSS)/(n - 1)}{URSS/((N(T - 1) - (k - 1))} \sim F_{n-1, n(T-1)-(k-1)},$$  \hspace{1cm} (5)$$

where $RRSS$ is the restricted residual sums of squares, coming from an OLS pool estimation, and $URSS$ is the unrestricted residual sums of squares, from the fixed effects estimation.

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

```matlab
>> regfe = panel(y, X, T, 'fe');
>> regfe.dvarnames = dvarnames;
>> regfe.ivarnames = ivarnames;
>> estprint(regfe);
```

Panel: Fixed effects (Within)

N observations: 816
N groups: 48
Obs per group: 17
R-squared = 0.941336
Adj R-squared = 0.941046
Joint significance: $F(4, 764) = 3064.808435$
  $p = 0.000$
Dept Var: lgsp

<table>
<thead>
<tr>
<th>Varname</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lpcap</td>
<td>-0.02615</td>
<td>0.02900</td>
<td>-0.9017</td>
<td>0.368</td>
</tr>
<tr>
<td>lpc</td>
<td>0.29201</td>
<td>0.02512</td>
<td>11.6246</td>
<td>0.000***</td>
</tr>
<tr>
<td>lemp</td>
<td>0.76816</td>
<td>0.03009</td>
<td>25.5273</td>
<td>0.000***</td>
</tr>
<tr>
<td>unemp</td>
<td>-0.00530</td>
<td>0.00099</td>
<td>-5.3582</td>
<td>0.000***</td>
</tr>
<tr>
<td>CONSTANT</td>
<td>2.35290</td>
<td>0.17481</td>
<td>13.4595</td>
<td>0.000***</td>
</tr>
</tbody>
</table>

Test of individual effects: $F(47, 764) = 75.820406$
  $p-value = 0.000$

The function `estprint` is used to display the table with the results taking the name of the variables specified in the properties `dvarnames` and `ivarnames` of the `estoutput` structure that is returned from the `panel` function.

---

3Where Wald is the standard Wald distance for joint significance tests of all estimated coefficients, excluding the constant term.
3.2. Between effects model

In the between estimation the parameters with the transformed variables:

\[
\hat{\beta}_{be} = (\bar{X}'\bar{X})^{-1}\bar{X}'\bar{y},
\]

where \(\bar{y}\) and \(\bar{X}\) are the means by groups, premultiplied by \(\sqrt{(T)}\) to take into account that the regression is based on \(nT\) observations, since the mean of each group is repeated \(T\) times, and it should be based on the \(n\) observations, Baltagi (2008). It is called “between” estimator because it takes into account the variation between groups, and since all observations are constant in each group. Again, this estimator is unbiased and consistent when \(n\) and \(T\) are large. Statistical inference is generally based on the asymptotic variance-covariance matrix:

\[
\text{VAR}(\hat{\beta}_{fe}) = S^2(\bar{X}'\bar{X})^{-1},
\]

where \(S^2\) denotes the residual variance: \(S^2 = (e^\top e)/(n-k)\), with residuals \(e = y - \bar{X}\hat{\beta}_{fe}\). The statistic of individual significance is distributed as a \(t\) student with \(n-k\) degrees of freedom. The Wald distance is computed as usual and the \(F\) statistic of joint significance is:

\[
F = \frac{\text{Wald}}{k-1} \sim F_{k-1,n-k},
\]

The goodness of fit is measured with the \(R^2\), which is computed as the square of the correlation coefficient of \(\bar{y}\) and \(\hat{\bar{y}}\).

The \texttt{panel} function implements the estimation of between effects panel data in MATLAB:

\begin{verbatim}
>> regbe = panel(y, X, T, 'be');
>> regbe.dvarnames = dvarnames;
>> regbe.ivarnames = ivarnames;
>> estprint(regbe);
\end{verbatim}

Panel: Between effects
N observations: 816
N groups: 48
Obs per group: 17
R-squared = 0.993909
Joint significance: \(F(4, 43) = 1754.114154\)
\(p = 0.0000\)
Dept Var: lgsp

\begin{tabular}{lccc}
Varname & Coefficient & Std. Error & Statistic & p-value \\
\hline
lpcap & 0.17937 & 0.07197 & 2.4922 & 0.017** \\
lpc & 0.30195 & 0.04182 & 7.2201 & 0.000*** \\
lemp & 0.57613 & 0.05637 & 10.2196 & 0.000*** \\
unemp & -0.00389 & 0.00991 & -0.3926 & 0.697 \\
CONSTANT & 1.58944 & 0.23298 & 6.8222 & 0.000*** \\
\end{tabular}
3.3. 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 \( i \)-th invariant time individual effects \( \mu_i \) and the disturbance \( v_{it} \).

\[
y_{it} = \alpha + X_{it}\beta + u_{it} \quad \forall i = 1, \ldots, n, \quad t = 1, \ldots, T
\]

(9)

The individual effect \( \mu_i \) is assumed independent of the disturbance \( v_{it} \). In addition, individual effects and disturbances are independent of the explanatory variables, i.e., \( \text{COV}(X_{it}, \mu_i) \neq 0 \) and \( \text{COV}(X_{it}, v_{it}) \neq 0 \) 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.

From the composed error component,

\[
E(\mu_i) = E(v_{it}) = E(\mu_i v_{it}) = 0
\]

(10)

\[
E(\mu_i \mu_j) = \begin{cases} \sigma^2_\mu & i \neq j \\ 0 & i = j \end{cases} \quad E(v_{it} v_{jt}) = \begin{cases} \sigma^2_v & i \neq j \\ 0 & i = j \end{cases}
\]

(11)

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

\[
\text{COV}(u_{it}, u_{js}) = \begin{cases} \sigma^2_\mu + \sigma^2_v & i = j, t = s \\ \sigma^2_\mu & i = j, t \neq s \end{cases}
\]

(12)

This implies the following correlation coefficient between disturbances:

\[
\rho = \text{CORR}(u_{it}, u_{js}) = \begin{cases} 1 & i = j, t = s \\ \sigma^2_\mu / (\sigma^2_\mu + \sigma^2_v) & i = j, t \neq s \end{cases}
\]

(13)

Therefore, the covariance matrix can be computed as follows:

\[
\Omega = E(uu^\top) = \sigma^2_\mu (I_n \otimes J_T) + \sigma^2_v (I_T \otimes I_T),
\]

(14)

where \( J_T \) is a matrix of ones of size \( T \) and the homoscedastic variance is \( \text{VAR}(u_{it}) = \sigma^2_\mu + \sigma^2_v \) for all \( i \) and \( t \). In this case, the GLS (generalized least squares) method yields an efficient estimator of the parameters. Following the general expression (White, 1980),

\[
\hat{\beta}_{re} = (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} y,
\]

(15)

with \( \Omega^{-1} = 1/\sigma^2_v [I_T - \sigma^2_\mu/\sigma^2_v + T \sigma^2_\mu] \). In order to obtain the GLS estimator of the regression coefficients, it is necessary to estimate \( \Omega^{-1} \) which is a matrix of dimension \( nT \times nT \). The GLS estimation of the random effects model is based on the transformation proposed by Baltagi (2008):\[
\hat{\beta}_{re} = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top y,
\]

(16)
where \( \tilde{y} = y - \theta \bar{y} \) and \( \tilde{X} = X - \theta \bar{X} \) are the transformed variables in quasideviations from the group mean. The factor \( \theta \) corresponds to Greene (2012):

\[
\theta = 1 - \sqrt{\frac{\sigma_v^2}{\sigma_v^2 + T\sigma_\mu^2}}. \tag{17}
\]

Focusing on a different derivation based on the spectral decomposition of \( \Omega \) one obtains, Baltagi (2008):

\[
\sigma_1^2 = T\sigma_\mu^2 + \sigma_v^2. \tag{18}
\]

The random effects estimator (16) is a weighted average of the within and between estimators, with the ratio \( \sigma_v^2 / (\sigma_v^2 + T\sigma_\mu^2) \) being the weight assigned to the between groups variation. Therefore, under the assumption of fixed effects this latter variation is omitted, with the ratio equal to zero and \( \theta \) equal to one (opposite to the OLS case). As a result, the treatment of individual effects as random provides an intermediate solution between complete variation and time invariant fixed effects.

Swamy and Arora (1972) suggest using the within regression residuals to compute \( \sigma_v^2 \) and the residuals from the between regression to compute \( \sigma_1^2 \). From these estimates \( \sigma_\mu^2 \) can be calculated as:

\[
\sigma_\mu^2 = \frac{\sigma_1^2 - \sigma_v^2}{T}. \tag{19}
\]

Statistical inference is generally based on the asymptotic variance-covariance matrix:

\[
\text{VAR}(\hat{\beta}_{re}) = S^2(\tilde{X}^\top \tilde{X})^{-1}, \tag{20}
\]

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

Finally, the statistic of individual significance is computed as usual and it is normally distributed. Also, Wald distance for joint significance is computed as before, and the statistic of joint significance is:

\[
\chi^2 = \text{Wald} \sim \chi_{k-1}^2. \tag{21}
\]

The goodness of fit is measured with the \( R^2 \), which is computed as the square of the correlation coefficient of \( \hat{y} \) and \( \tilde{y} \).

The \texttt{panel} function implements the estimation of random effects panel data in MATLAB:

\begin{verbatim}
>> regre = panel(y, X, T, 're');
>> regre.dvarnames = dvarnames;
>> regre.ivarnames = ivarnames;
>> estprint(regre);
\end{verbatim}

\footnote{If the estimated \( \sigma_\mu^2 \) is negative, which occurs when the true value is closed to zero (Baltagi 2008, p. 20), it may be replaced by zero as suggested by Maddala and Mount (1973).}
Panel: Random effects GLS (Swamy and Arora)

N observations: 816  
N groups: 48  
Obs per group: 17  
R-squared = 0.959332  
Joint significance: Chi2(4) = 19131.085009  
p-value = 0.0000  

Dept Var: lgsp

<table>
<thead>
<tr>
<th>Varname</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lpcap</td>
<td>0.00444</td>
<td>0.02342</td>
<td>0.1895</td>
<td>0.850</td>
</tr>
<tr>
<td>lpc</td>
<td>0.31055</td>
<td>0.01980</td>
<td>15.6805</td>
<td>0.000***</td>
</tr>
<tr>
<td>lemp</td>
<td>0.72967</td>
<td>0.02492</td>
<td>29.2803</td>
<td>0.000***</td>
</tr>
<tr>
<td>unemp</td>
<td>-0.00617</td>
<td>0.00091</td>
<td>-6.8033</td>
<td>0.000***</td>
</tr>
<tr>
<td>CONSTANT</td>
<td>2.13541</td>
<td>0.13346</td>
<td>16.0002</td>
<td>0.000***</td>
</tr>
</tbody>
</table>

sigma_mu = 0.082691  
sigma_v = 0.038137  
sigma_l = 0.343068  
Theta = 0.888835

Here \( \rho_{\mu} \) is the fraction of variance due to the individual effects and it is computed as 
\( \rho_{\mu} = \sigma_{\mu}^2 / (\sigma_{\mu}^2 + \sigma_v^2) \).

3.4. Hausman test of specification

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 individuals effects and the explanatory variables are correlated: \( E(\mu_iX_{it}) \neq 0 \), the fixed effects model provides an unbiased estimator, otherwise a feasible GLS is an efficient estimator in a random effects model.

Hausman (1978) suggests comparing 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 \( H_0 : E(\mu_iX_{it}) = 0 \). Under \( H_0 \) the GLS estimator is BLUE, consistent and asymptotically efficient, while the within estimator is consistent whether \( H_0 \) is true or not. Furthermore, the GLS estimator is inconsistent if \( H_0 \) is false. Therefore, the statistic would be based on the difference between both estimators: \( \hat{\beta}_{fe} - \hat{\beta}_{re} \).

Hence, the Hausman test statistic is given by (Baltagi (2008)):

\[
H = (\hat{\beta}_{fe} - \hat{\beta}_{re})^T \text{VAR}(\hat{\beta}_{fe} - \hat{\beta}_{re})^{-1}(\hat{\beta}_{fe} - \hat{\beta}_{re}) \sim \chi^2_{k-1},
\]  

(22)

where \( \text{VAR}(\hat{\beta}_{fe} - \hat{\beta}_{re})^{-1} = \text{VAR}(\hat{\beta}_{fe}) - \text{VAR}(\hat{\beta}_{re}) \).

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 \([H, p] = \text{hausman}(\text{estA}, \text{estB})\) function implements the Hausman test in MATLAB, where the input arguments \text{estA} and \text{estB} are \text{estoutput} structures of the previous estimations. The function returns the value of the test, \( H \), and its corresponding \( p \) value, \( p \). To display the results in a table, the \text{hausmanprint(estA, estB)} must be used:
Hausman’s test of specification

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>lpcap</td>
<td>-0.026150</td>
<td>0.004439</td>
<td>-0.030588</td>
<td>0.01711</td>
</tr>
<tr>
<td>lpc</td>
<td>0.292007</td>
<td>0.310548</td>
<td>-0.018542</td>
<td>0.01545</td>
</tr>
<tr>
<td>lemp</td>
<td>0.768159</td>
<td>0.729671</td>
<td>0.038489</td>
<td>0.01687</td>
</tr>
<tr>
<td>unemp</td>
<td>-0.005298</td>
<td>-0.006172</td>
<td>0.000875</td>
<td>0.00039</td>
</tr>
</tbody>
</table>

estA is consistent under H0 and H1 (estA = Panel-fe)
estB is consistent under H0 (estB = Panel-re)
H0: coef(estA) - coef(estB) = 0
H1: coef(estA) - coef(estB) != 0
H = 9.5254156
p-value = 0.04923
H is distributed as Chi2(4)

3.5. Heteroscedasticity in panel data models

The fixed effects model can be estimated using the within estimator and a robust covariance matrix when the disturbances are affected by heteroscedasticity. Hansen (2007) proposed a robust estimation of the parameters’ covariance matrix using the White sandwich estimator, White (1980):

\[
\text{VAR}(\hat{\beta}_{fe}) = \frac{n}{n-1} \left( \frac{N - 1}{N - k} (\bar{X}^\top \bar{X})^{-1} \left[ \sum_{i=1}^{n} \bar{X}_i e_i e_i^\top \bar{X}_i \right] (\bar{X}^\top \bar{X})^{-1} \right)
\]  \tag{23}

From \text{VAR}(\hat{\beta}_{fe}) the correct variance of the constant term must be computed as:

\[
\text{VAR}(\hat{\alpha}_{fe}) = \left( \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{T} \sum_{t=1}^{T} X_{it} \right) \right) \text{VAR}(\hat{\beta}_{fe}) \left( \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{T} \sum_{t=1}^{T} X_{it} \right) \right)^\top
\]  \tag{24}

For the random effects model the robust estimation of the parameters’ covariance matrix is computed using an estimator equivalent to that proposed by White (1980), (23), but with the suitable transformation of the variables.
The \text{panel} function, with the \textbf{options} argument set to \textit{robust}, implements the estimation of fixed effects robust panel data models in MATLAB:

\[
\text{>> regfer = panel(y, X, T, 'fe', 'robust');}
\text{>> regfer.dvarnames = dvarnames;}
\text{>> regfer.ivarnames = ivarnames;}
\text{>> estprint(regfer);}
\]
Panel: Fixed effects (Within)

N observations: 816
N groups: 48
Obs per group: 17
R-squared = 0.941336
Adj R-squared = 0.941046
Joint significance: F(4, 47) = 395.610133
   p = 0.0000
Dept Var: lgsp
Robust standard errors adjusted for 48 clusters

<table>
<thead>
<tr>
<th>Varname</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lpcap</td>
<td>-0.02615</td>
<td>0.06111</td>
<td>-0.4279</td>
<td>0.671</td>
</tr>
<tr>
<td>lpc</td>
<td>0.29221</td>
<td>0.06255</td>
<td>4.6684</td>
<td>0.000***</td>
</tr>
<tr>
<td>lemp</td>
<td>0.76816</td>
<td>0.08273</td>
<td>9.2848</td>
<td>0.000***</td>
</tr>
<tr>
<td>unemp</td>
<td>-0.00530</td>
<td>0.00253</td>
<td>-2.0952</td>
<td>0.042**</td>
</tr>
<tr>
<td>CONSTANT</td>
<td>2.35290</td>
<td>0.31459</td>
<td>7.4792</td>
<td>0.000***</td>
</tr>
</tbody>
</table>

In the random effects model the robust option provides the robust covariance matrix estimation:

```matlab
>> regrer = panel(y, X, T, 're', 'robust');
>> regrer.dvarnames = dvarnames;
>> regrer.ivarnames = ivarnames;
>> estprint(regrer);
```

Panel: Random effects GLS (Swamy and Arora)

N observations: 816
N groups: 48
Obs per group: 17
R-squared = 0.959332
Joint significance: Chi2(4) = 4408.644223
   p-value = 0.0000
Dept Var: lgsp
Robust standard errors adjusted for 48 clusters

<table>
<thead>
<tr>
<th>Varname</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lpcap</td>
<td>0.00444</td>
<td>0.05531</td>
<td>0.0802</td>
<td>0.936</td>
</tr>
<tr>
<td>lpc</td>
<td>0.31055</td>
<td>0.04416</td>
<td>7.0320</td>
<td>0.000***</td>
</tr>
<tr>
<td>lemp</td>
<td>0.72967</td>
<td>0.07088</td>
<td>10.2941</td>
<td>0.000***</td>
</tr>
<tr>
<td>unemp</td>
<td>-0.00617</td>
<td>0.00236</td>
<td>-2.6120</td>
<td>0.009***</td>
</tr>
<tr>
<td>CONSTANT</td>
<td>2.13541</td>
<td>0.24179</td>
<td>8.8318</td>
<td>0.000***</td>
</tr>
</tbody>
</table>

sigma_mu = 0.082691
sigma_v = 0.038137
sigma_I = 0.343068
rho_mu = 0.824601
Theta = 0.888835
4. Instrumental panel data models

The assumption of exogeneity of the independent variables, $X$, when they are uncorrelated with the disturbance, $E(X_t, u_t) = 0$, implies that OLS remains valid. However, there are many applications in which this assumption is untenable. In this case, when the regressors are endogenous, the OLS estimator loses consistency and unbiasedness. Consequently, we can apply an instrumental variables (IV) two stage estimation to the fixed effects, random effects and between models, Greene (2012).

We assume that there is a set of variables that are exogenous, uncorrelated with the disturbance, and relevant, i.e., correlated with the endogenous independent variables. This set is represented by the $H$ matrix.

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.\(^5\) We estimate the consumption, $c$, measured as per capita sales, which depends on the price per pack, $\text{price}$, per capita disposable income, $\text{ndi}$, and the minimum price in neighbor states, $\text{pimin}$. The instruments normally used are the lags of the disposable income, $\text{ndi}_1$, and the lag of the minimum price $\text{pimin}_1$.\(^6\)

```matlab
>> load('CigarData.mat')
>> y = log(c);
>> X = [log(price), log(ndi), log(pimin)];
>> H = [log(ndi_1), log(pimin_1), log(ndi), log(pimin)];
>> T = 29;
>> dvarnames = {'lc'};
>> ivarnames = {'lprice', 'lndi', 'lpimin'};
```

Instrumental panel data models are estimated using the `ivpanel(y, X, H, T, method)` function, where $y$ is the vector of the dependent variable, $X$ is the matrix of explanatory variables, $H$ is the matrix of instruments, $T$ is the number of time periods per unit, and $\text{method}$ is a string that specifies the choice of panel data estimation method, among the following:

- $\hat{\text{po}}$: for a pool estimation.
- $\hat{\text{fe}}$: for a fixed effects (within) estimation.
- $\hat{\text{be}}$: for a between effects estimation.
- $\hat{\text{re}}$: for a random effects estimation
- $\hat{\text{ec}}$: for an error-components estimation, Baltagi and Liu (2009).

\(^5\)Data is available in MATLAB format in the supplementary file `CigarData.mat`.

\(^6\)The equation we estimate differs from the original one, which corresponds to a dynamic panel data model.
4.1. Two stage least squares (2SLS)

The first stage of the 2SLS estimation consists of estimating the independent variables, \( \hat{X} \), by an OLS estimate of \( X \) over the exogenous variables and instruments, \( H \):

\[
\hat{X} = \hat{H}(\hat{H}^{-} \hat{H})^{-1}\hat{H}^{-} \hat{X}
\]  

(25)

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

\[
\hat{\beta}_{2SLS} = (\hat{X}^{-} \hat{X})^{-1}\hat{X}^{-} \hat{y}
\]  

(26)

In each case, \( \hat{y} \), \( \hat{X} \) and \( \hat{H} \) represent the different transformations applied to the variables to obtain the within, between and GLS estimator as explained in Section 3. Regarding statistical inference, the statistic of individual significance is normally distributed, while the statistic of joint significance is distributed as a \( \chi^2 \) with \( k - 1 \) degrees of freedom. The test for individual effects is that proposed in Baltagi (2008).

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

```matlab
>> regivfe = ivpanel(y, X, H, T, 'fe');
>> regivfe.dvarnames = dvarnames;
>> regivfe.ivarnames = ivarnames;
>> estprint(regivfe)
```

**IV Panel: Fixed effects (Within)**

- N observations: 1334
- N groups: 46
- Obs per group: 29
- R-squared = 0.640642
- Joint significance: Chi2(3) = 1792.756633
  - p-value = 0.0000

Dept Var: lc

<table>
<thead>
<tr>
<th>Varname</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lprice</td>
<td>-1.01636</td>
<td>0.24920</td>
<td>-4.0785</td>
<td>0.000***</td>
</tr>
<tr>
<td>lndi</td>
<td>0.53785</td>
<td>0.02303</td>
<td>23.3507</td>
<td>0.000***</td>
</tr>
<tr>
<td>lpimin</td>
<td>0.31237</td>
<td>0.22839</td>
<td>1.3677</td>
<td>0.171</td>
</tr>
<tr>
<td>CONSTANT</td>
<td>2.99141</td>
<td>0.08111</td>
<td>36.8827</td>
<td>0.000***</td>
</tr>
</tbody>
</table>

```matlab
>> regivbe = ivpanel(y, X, H, T, 'be');
>> regivbe.dvarnames = dvarnames;
>> regivbe.ivarnames = ivarnames;
>> estprint(regivbe)
```

Note that the matrix \( H \) must include the instruments as well as the exogenous variables that are also included in \( X \), which are instruments of themselves.
IV Panel: Between effects

N observations: 1334
N groups: 46
Obs per group: 29
R-squared = 0.311151
Joint significance: Chi2(3) = 6.660389
p-value = 0.0835

Dept Var: lc

<table>
<thead>
<tr>
<th>Varname</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lprice</td>
<td>-3.27523</td>
<td>2.61392</td>
<td>-1.2530</td>
<td>0.210</td>
</tr>
<tr>
<td>lndi</td>
<td>0.83220</td>
<td>0.40039</td>
<td>2.0785</td>
<td>0.038**</td>
</tr>
<tr>
<td>lpimin</td>
<td>1.18107</td>
<td>1.32375</td>
<td>0.8922</td>
<td>0.372</td>
</tr>
<tr>
<td>CONSTANT</td>
<td>6.17390</td>
<td>3.29673</td>
<td>1.8727</td>
<td>0.061*</td>
</tr>
</tbody>
</table>

```matlab
>> regivre = ivpanel(y, X, H, T, 're');
>> regivre.dvarnames = dvarnames;
>> regivre.ivarnames = ivarnames;
>> estprint(regivre)
```

IV Panel: Random effects GLS (Swamy and Arora)

N observations: 1334
N groups: 46
Obs per group: 29
R-squared = 0.638272
Joint significance: Chi2(3) = 1820.426405
p-value = 0.0000

Dept Var: lc

<table>
<thead>
<tr>
<th>Varname</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lprice</td>
<td>-1.00711</td>
<td>0.24735</td>
<td>-4.0715</td>
<td>0.000***</td>
</tr>
<tr>
<td>lndi</td>
<td>0.53747</td>
<td>0.02303</td>
<td>23.3398</td>
<td>0.000***</td>
</tr>
<tr>
<td>lpimin</td>
<td>0.30357</td>
<td>0.22643</td>
<td>1.3407</td>
<td>0.180</td>
</tr>
<tr>
<td>CONSTANT</td>
<td>2.99212</td>
<td>0.08567</td>
<td>34.9268</td>
<td>0.000***</td>
</tr>
</tbody>
</table>

σμ = 0.190101
σv = 0.077566
σ1 = 1.026661
ρμ = 0.857278
Θ = 0.924449
4.2. Error components two stage least squares (EC2SLS)

Baltagi (1981) and Baltagi and Liu (2009) propose a generalized two stage least squares (G2SLS) estimation using the following matrix of instruments:

\[ A = \begin{bmatrix} \tilde{H}, \bar{H} \end{bmatrix}, \quad (27) \]

where \(\tilde{H}\) contains the transformed instruments in deviations from the group mean, and \(\bar{H}\) the group means. The 2SLS estimation is then performed using this matrix of instruments.\(^8\)

The error components two stage least squares (EC2SLS) estimator is consistent and presents the same limiting distribution than the G2SLS estimator. Although it is worth noting that for small samples the former shows gains in efficiency, Baltagi and Liu (2009).

The `ivpanel` function provides an estimation of the error components two stage least squares (EC2SLS) model in MATLAB by specifying the `ec` method:

\[
\begin{align*}
&\text{regivec = ivpanel(y, X, H, T, 'ec');} \\
&\text{regivec.dvarnames = dvarnames;} \\
&\text{regivec.ivarnames = ivarnames;} \\
&\text{estprint(regivec)}
\end{align*}
\]

IV Panel: Error components (EC2SLS)

N observations: 1334  
N groups: 46  
Obs per group: 29  
R-squared = 0.638782  
Joint significance: Chi2(3) = 1825.252894  
p-value = 0.0000

Dept Var: lc

\[
\begin{array}{cccc}
\text{Varname} & \text{Coefficient} & \text{Std. Error} & \text{Statistic} & \text{p-value} \\
\hline
\text{lprice} & -0.99268 & 0.23587 & -4.2086 & 0.000*** \\
\text{lndi} & 0.53641 & 0.02236 & 23.9939 & 0.000*** \\
\text{lpimin} & 0.29039 & 0.21597 & 1.3446 & 0.179 \\
\text{CONSTANT} & 2.99512 & 0.08420 & 35.5724 & 0.000*** \\
\end{array}
\]

\[
\begin{align*}
\text{sigma_mu} & = 0.190101 \\
\text{sigma_v} & = 0.077566 \\
\text{sigma_1} & = 1.026661 \\
\text{rho_mu} & = 0.857278 \\
\text{Theta} & = 0.924449
\end{align*}
\]

5. Spatial panel data models

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 followed over time. In general, a spatial panel data set contains more information and less multicollinearity among the variables

\[^8\text{The instruments } A \text{ are used in the 2SLS procedure, but only } H \text{ is used when estimating } \sigma_v^2 \text{ and } \sigma_1^2.\]
than a cross-section spatial counterpart (see Anselin (1988, 2010) for an introduction to this literature). Additionally, the use of panel data increases the efficiency due to larger degrees of freedom and allows the inclusion of unobservable heterogeneities Baltagi (2008).

In the context of cross-sectional models, Kelejian and Prucha (1998) introduced a generalized spatial two-stage least squares estimator, Kelejian and Prucha (1999)\(^9\) proposed a generalized moments (GM) estimation method feasible even when \(n\) is large, while Anselin (1988) provided the ML (Maximum likelihood) estimator. Kapoor, Kelejian, and Prucha (2007) generalized the GM procedure from cross-section to panel data and derived its properties when \(T\) is fixed and \(n\) tends to infinite. Most recently, Elhorst (2003, 2010) and Lee and Yu (2010) presented 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. In line with Anselin (1988) and Kapoor et al. (2007), Baltagi, Egger, and Pfaffermayr (2006) suggest a generalized spatial panel model allowing for spatial correlation in the individual and the remainder error components. They derive the ML estimator for this more general spatial panel model with random effects.

In order to compute different estimators in spatial panel models, we consider the Cliff-Ord autoregressive spatial panel model:

\[
y_{it} = \lambda W y_{it} + \beta X + \beta L W L + u_i + v_{it},
\]

where the matrix \(L\) contains the spatial lagged independent variables, which usually are also included in \(X\).

The application is based on Munnell (1990) and Baltagi (2008) data of U.S. states production as in Section 3.\(^10\)

\[
W_{\text{big}} = W \otimes I_T
\]


\(^{10}\)Munnel (1980) data is available in MATLAB format in the supplementary file MunnellData.mat, and the \(W\) matrix in the file MunnellW.mat.
Spatial panel data models are estimated using the `spanel(y, X, L, W, T, method)` function, where `y` is the vector of the dependent variable, `X` is the matrix of explanatory variables, `L` is the matrix of spatial lagged independent variables, `T` is the number of time periods per unit, and `method` is a string that specifies the panel data estimation method to use, among the following:

- **po**: for a spatial pool estimation.
- **fe**: for a spatial fixed effects (within) estimation.
- **be**: for a spatial between effects estimation.
- **re**: for a spatial random effects estimation.
- **ec**: for a spatial error components estimation, Baltagi and Liu (2011).
- **sec-b**: for a spatial error components best estimation, Baltagi and Liu (2011).

### 5.1. Generalized two stage least squares (GS2SLS)

The spatial panels are computed as an instrumental variable estimation, extending the generalized spatial two stage least squares estimator (GS2SLS) provided by Kelejian and Prucha (1998) with fixed, between and random effects.

For simplicity, we rewrite the model more compactly as follows:

\[ y_{it} = \delta Z_{it} + u_{it}, \quad (30) \]

where \( Z_{it} = (W y_{it}, X_{it}, W L_{it}) \) and \( \delta = (\lambda, \beta_X, \beta_L) \). Following Kelejian and Prucha (1998) we build the matrix of instruments as:

\[ H = [X, WX, W^2 X] \quad (31) \]

We compute the first stage of the GS2SLS method estimating the fitted values for the independent variables \( \hat{Z} \) performing OLS of \( Z \) on the instruments \( H \):

\[ \hat{Z} = \bar{H}(\bar{H}^\top \bar{H})^{-1}\bar{H}^\top \bar{Z}. \quad (32) \]

In the second stage we compute the coefficients, \( \hat{\delta} \), using the predicted \( \hat{Z} \):

\[ \hat{\delta} = (\hat{Z}^\top \hat{Z})^{-1}\hat{Z}^\top \hat{y} \quad (33) \]

In each case, \( \hat{y} \), \( \hat{X} \) and \( \hat{Z} \) represent the different transformations applied to their corresponding set of variables to obtain the alternative estimations: fixed effects spatial two stage least squares (FE-S2SLS), between effects spatial two stage least squares (BE-2SLS), and random effects spatial two stage least squares (RE-S2SLS).

The fitted values are computed as in Elhorst (2003, 2010):

\[ \hat{y} = (I_N - \lambda W)^{-1}(X \beta_X + WL \beta_L) \quad (34) \]
The `spanel` function implements the estimation of the fixed, between, random and error components spatial panel data models in MATLAB:

\[
\text{\texttt{>> regsfe = spanel(y, X, L, W, T, 'fe');}}
\]
\[
\text{\texttt{>> regsfe.dvarnames = dvarnames;}}
\]
\[
\text{\texttt{>> regsfe.ivarnames = ivarnames;}}
\]
\[
\text{\texttt{>> estprint(regsfe);}}
\]

**Spatial Panel: Fixed effects (FE-2SLS)**

- N observations: 816
- N groups: 48
- Obs per group: 17
- Joint significance: Chi2(5) = 12845.284388
  - p-value = 0.0000

<table>
<thead>
<tr>
<th>Varname</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lpcap</td>
<td>-0.04041</td>
<td>0.02846</td>
<td>-1.4197</td>
<td>0.156</td>
</tr>
<tr>
<td>lpc</td>
<td>0.21904</td>
<td>0.02679</td>
<td>8.1770</td>
<td>0.000***</td>
</tr>
<tr>
<td>lemp</td>
<td>0.66833</td>
<td>0.03285</td>
<td>20.3447</td>
<td>0.000***</td>
</tr>
<tr>
<td>unemp</td>
<td>-0.00473</td>
<td>0.00097</td>
<td>-4.8683</td>
<td>0.000***</td>
</tr>
<tr>
<td>W*lgsp</td>
<td>0.19166</td>
<td>0.02794</td>
<td>6.8597</td>
<td>0.000***</td>
</tr>
<tr>
<td>CONSTANT</td>
<td>1.93735</td>
<td>0.18150</td>
<td>10.6741</td>
<td>0.000***</td>
</tr>
</tbody>
</table>

\[
\text{\texttt{>> regsbe = spanel(y, X, L, W, T, 'be');}}
\]
\[
\text{\texttt{>> regsbe.dvarnames = dvarnames;}}
\]
\[
\text{\texttt{>> regsbe.ivarnames = ivarnames;}}
\]
\[
\text{\texttt{>> estprint(regsbe);}}
\]

**Spatial Panel: Between effects (BE-2SLS)**

- N observations: 816
- N groups: 48
- Obs per group: 17
- Joint significance: Chi2(5) = 6901.939058
  - p-value = 0.0000

<table>
<thead>
<tr>
<th>Varname</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lpcap</td>
<td>0.17131</td>
<td>0.07487</td>
<td>2.2882</td>
<td>0.022**</td>
</tr>
<tr>
<td>lpc</td>
<td>0.30163</td>
<td>0.04217</td>
<td>7.1520</td>
<td>0.000***</td>
</tr>
<tr>
<td>lemp</td>
<td>0.58559</td>
<td>0.06082</td>
<td>9.6283</td>
<td>0.000***</td>
</tr>
<tr>
<td>unemp</td>
<td>-0.00242</td>
<td>0.01054</td>
<td>-0.2297</td>
<td>0.818</td>
</tr>
<tr>
<td>W*lgsp</td>
<td>-0.01082</td>
<td>0.02474</td>
<td>-0.4373</td>
<td>0.662</td>
</tr>
<tr>
<td>CONSTANT</td>
<td>1.70896</td>
<td>0.36036</td>
<td>4.7423</td>
<td>0.000***</td>
</tr>
</tbody>
</table>
A Panel Data Toolbox for MATLAB

>> regsre = spanel(y, X, L, W, T, 're');
>> regsre.dvarnames = dvarnames;
>> regsre.ivarnames = ivarnames;
>> estprint(regsre);

Spatial Panel: Random effects (RE-2SLS)

N observations: 816
N groups: 48
Obs per group: 17
Joint significance: Chi2(5) = 18847.914957
  p-value = 0.0000
Dept Var: lgsp

<table>
<thead>
<tr>
<th>Varname</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lpcap</td>
<td>0.02286</td>
<td>0.02492</td>
<td>0.9173</td>
<td>0.359</td>
</tr>
<tr>
<td>lpc</td>
<td>0.29376</td>
<td>0.02104</td>
<td>13.9596</td>
<td>0.000***</td>
</tr>
<tr>
<td>lemp</td>
<td>0.70864</td>
<td>0.02679</td>
<td>26.4514</td>
<td>0.000***</td>
</tr>
<tr>
<td>unemp</td>
<td>-0.00648</td>
<td>0.00092</td>
<td>-7.0525</td>
<td>0.000***</td>
</tr>
<tr>
<td>W*lgsp</td>
<td>0.03547</td>
<td>0.01481</td>
<td>2.3946</td>
<td>0.017**</td>
</tr>
<tr>
<td>CONSTANT</td>
<td>1.90996</td>
<td>0.16628</td>
<td>11.4865</td>
<td>0.000***</td>
</tr>
</tbody>
</table>

sigma_mu = 0.083405
sigma_v  = 0.037325
sigma_l  = 0.345907
rho_mu   = 0.833143
Theta    = 0.892094

>> regsec = spanel(y, X, L, W, T, 'ec');
>> regsec.dvarnames = dvarnames;
>> regsec.ivarnames = ivarnames;
>> estprint(regsec);

Spatial Panel: Error Components (SEC-2SLS)

N observations: 816
N groups: 48
Obs per group: 17
Joint significance: Chi2(5) = 18842.897203
  p-value = 0.0000
Dept Var: lgsp

<table>
<thead>
<tr>
<th>Varname</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lpcap</td>
<td>0.02454</td>
<td>0.02491</td>
<td>0.9850</td>
<td>0.325</td>
</tr>
<tr>
<td>lpc</td>
<td>0.29239</td>
<td>0.02104</td>
<td>13.8978</td>
<td>0.000***</td>
</tr>
<tr>
<td>lemp</td>
<td>0.70670</td>
<td>0.02678</td>
<td>26.3879</td>
<td>0.000***</td>
</tr>
<tr>
<td>unemp</td>
<td>-0.00651</td>
<td>0.00092</td>
<td>-7.0840</td>
<td>0.000***</td>
</tr>
<tr>
<td>W*lgsp</td>
<td>0.03850</td>
<td>0.01476</td>
<td>2.6089</td>
<td>0.009***</td>
</tr>
<tr>
<td>CONSTANT</td>
<td>1.89001</td>
<td>0.16608</td>
<td>11.3803</td>
<td>0.000***</td>
</tr>
</tbody>
</table>

sigma_mu = 0.083405
sigma_v  = 0.037325
sigma_l  = 0.345907
rho_mu   = 0.833143
Theta    = 0.892094
5.2. Spatial error components best two stage least squares (SEC-B2SLS)

Baltagi and Liu (2011) extend the error component two-stage least square estimator proposed by Baltagi (1981), following the method introduced by Kelejian and Prucha (1998) and using Lee (2003) optimal instrument for this spatial autoregressive panel model. They obtain the spatial error components best two stage least squares estimator (SEC-B2SLS), in which we base our estimation.

Accordingly, we consider the following matrix of instruments:

\[
B = [\tilde{H}_b^*, \bar{H}_b^*],
\]

where \(\tilde{H}_b^* = [\bar{X}, WA^{-1} \bar{X} \hat{\beta}]\) and \(\bar{H}_b^* = [\bar{X}, WA^{-1} \bar{X} \hat{\beta}]\) are the instruments with the transformations used in the fixed and between models, respectively, and \(A = (I_N - \lambda W)\). \(\lambda\) and \(\beta\) are consistent estimators and can be those obtained from a pool spatial regression. Then, GSL estimation is performed using the matrix of instruments \(B\).

```matlab
>> regsecb = spanel(y, X, L, W, T, 'sec−b');
>> regsecb.dvarnames = dvarnames;
>> regsecb.ivarnames = ivarnames;
>> estprint(regsecb);
```

Spatial Panel: Error Components Best (SEC−B2SLS)

N observations: 816
N groups: 48
Obs per group: 17
R-squared = 0.958805
Joint significance: Chi2(5) = 18852.569351
p-value = 0.0000
Dept Var: lgsp

<table>
<thead>
<tr>
<th>Varname</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lpcap</td>
<td>0.01838</td>
<td>0.02493</td>
<td>0.7371</td>
<td>0.461</td>
</tr>
<tr>
<td>lpc</td>
<td>0.29742</td>
<td>0.02105</td>
<td>14.1284</td>
<td>0.000***</td>
</tr>
<tr>
<td>lemp</td>
<td>0.71380</td>
<td>0.02680</td>
<td>26.6310</td>
<td>0.000***</td>
</tr>
<tr>
<td>unemp</td>
<td>-0.00640</td>
<td>0.00092</td>
<td>-6.9677</td>
<td>0.000***</td>
</tr>
<tr>
<td>W*lgsp</td>
<td>0.02739</td>
<td>0.01489</td>
<td>1.8396</td>
<td>0.066*</td>
</tr>
<tr>
<td>CONSTANT</td>
<td>1.96319</td>
<td>0.16656</td>
<td>11.7869</td>
<td>0.000***</td>
</tr>
</tbody>
</table>

\(\sigma_\mu = 0.083405\)
\(\sigma_v = 0.037325\)
\(\sigma_1 = 0.345907\)
\(\rho_\mu = 0.833143\)
\(\Theta = 0.892094\)

6. Numerical checks

Numerical checks against other commercial and free software are performed by comparing the standard panel data results obtained in Section 3 from this Panel Data Toolbox in MATLAB and the results reported by Stata, `xtreg` function, and the R package `plm` by Croissant and Millo (2008), `plm` function.
Results for the fixed, between and random estimators using the Munnell (1990) data are reported in Table 1. The decimal places are those corresponding to the default output of all three softwares. Results show that there are not differences in the estimated coefficients and \( t \)-statistics between the three programs.

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th></th>
<th></th>
<th>Coefficient</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MATLAB</td>
<td>Stata</td>
<td>R</td>
<td>MATLAB</td>
<td>Stata</td>
<td>R</td>
</tr>
<tr>
<td>Fixed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( lpcap )</td>
<td>-0.02615</td>
<td>-0.0261493</td>
<td>-0.02614965</td>
<td>-0.9017</td>
<td>-0.90</td>
<td>-0.9017</td>
</tr>
<tr>
<td>( lpc )</td>
<td>0.29201</td>
<td>0.2920067</td>
<td>0.29200693</td>
<td>11.6246</td>
<td>11.62</td>
<td>11.6246</td>
</tr>
<tr>
<td>( lemp )</td>
<td>0.76816</td>
<td>0.7681595</td>
<td>0.76815947</td>
<td>25.5273</td>
<td>25.53</td>
<td>25.5273</td>
</tr>
<tr>
<td>( unemp )</td>
<td>-0.00530</td>
<td>-0.0052977</td>
<td>-0.00529774</td>
<td>-5.3582</td>
<td>-5.36</td>
<td>-5.3582</td>
</tr>
<tr>
<td>( CONST )</td>
<td>2.35290</td>
<td>2.352898</td>
<td>N.A.</td>
<td>13.4595</td>
<td>13.46</td>
<td>N.A.</td>
</tr>
<tr>
<td>Between</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( lpcap )</td>
<td>0.17937</td>
<td>0.1793651</td>
<td>0.1793651</td>
<td>2.4922</td>
<td>2.49</td>
<td>2.4922</td>
</tr>
<tr>
<td>( lpc )</td>
<td>0.30195</td>
<td>0.3019542</td>
<td>0.3019542</td>
<td>7.2201</td>
<td>7.22</td>
<td>7.2201</td>
</tr>
<tr>
<td>( lemp )</td>
<td>0.57613</td>
<td>0.5761274</td>
<td>0.5761274</td>
<td>10.2196</td>
<td>10.22</td>
<td>10.2196</td>
</tr>
<tr>
<td>( unemp )</td>
<td>-0.00389</td>
<td>-0.0038903</td>
<td>-0.0038903</td>
<td>-0.3926</td>
<td>-0.39</td>
<td>-0.3926</td>
</tr>
<tr>
<td>( CONST )</td>
<td>1.58944</td>
<td>1.589444</td>
<td>1.5894444</td>
<td>6.8222</td>
<td>6.82</td>
<td>6.8222</td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( lpcap )</td>
<td>0.00444</td>
<td>0.0044388</td>
<td>0.00443859</td>
<td>0.1895</td>
<td>0.19</td>
<td>0.1895</td>
</tr>
<tr>
<td>( lpc )</td>
<td>0.31055</td>
<td>0.3105483</td>
<td>0.31054843</td>
<td>15.6805</td>
<td>15.68</td>
<td>15.6805</td>
</tr>
<tr>
<td>( lemp )</td>
<td>0.72967</td>
<td>0.7296705</td>
<td>0.72967053</td>
<td>29.2803</td>
<td>29.28</td>
<td>29.2803</td>
</tr>
<tr>
<td>( unemp )</td>
<td>-0.00617</td>
<td>-0.0061725</td>
<td>-0.00617247</td>
<td>-6.8033</td>
<td>-6.80</td>
<td>-6.8033</td>
</tr>
<tr>
<td>( CONST )</td>
<td>2.13541</td>
<td>2.135411</td>
<td>2.13541100</td>
<td>16.0002</td>
<td>16.00</td>
<td>16.0002</td>
</tr>
</tbody>
</table>

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

Checks for the instrumental variables panel data models with fixed, between, random, and error components for Stata, using the `xtivreg` function, and R package `plm` function by Croissant and Millo (2008), are reported in Table 2, using the cigarette data, Baltagi (2008). Again, results are the the same for all three programs.

Spatial panel estimations are checked against the R package `splm` by Millo and Piras (2012), using the `spgm` function, which performs a GM implementation. Results in Table 3 reveal slight differences in the estimated coefficients and \( t \) statistics, but these differences do not change the overall features of the estimation results.

---

11 All decimals can be obtained for the Panel Data Toolbox accessing directly the properties `coef` or `statistic` of the `estoutput` structure.

12 The code is available in the supplementary files `NC_panel_Stata.do` and `NC_panel_R.R`.

13 The code is available in the supplementary files `NC_ivpanel_Stata.do` and `NC_ivpanel_R.R`.

14 The R package `sphet` by Piras (2010) can estimate spatial models with heteroskedastic innovations.

15 The code is available in the supplementary file `NC_spanel_R.R`.
Table 2: Comparison of estimated coefficients and *t* statistics for instrumental panel data against Stata and R.

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th><em>t</em> statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MATLAB</td>
<td>Stata</td>
</tr>
<tr>
<td>Fixed</td>
<td>lprice</td>
<td>-1.01636</td>
</tr>
<tr>
<td></td>
<td>lndi</td>
<td>0.53785</td>
</tr>
<tr>
<td></td>
<td>lpimin</td>
<td>0.31237</td>
</tr>
<tr>
<td></td>
<td>CONST</td>
<td>2.99141</td>
</tr>
<tr>
<td>Between</td>
<td>lprice</td>
<td>-3.27523</td>
</tr>
<tr>
<td></td>
<td>lndi</td>
<td>0.83220</td>
</tr>
<tr>
<td></td>
<td>lpimin</td>
<td>1.18107</td>
</tr>
<tr>
<td></td>
<td>CONST</td>
<td>6.17390</td>
</tr>
<tr>
<td>Random</td>
<td>lprice</td>
<td>-1.00711</td>
</tr>
<tr>
<td></td>
<td>lndi</td>
<td>0.53747</td>
</tr>
<tr>
<td></td>
<td>lpimin</td>
<td>0.30357</td>
</tr>
<tr>
<td></td>
<td>CONST</td>
<td>2.99212</td>
</tr>
<tr>
<td>Error components</td>
<td>lprice</td>
<td>-0.99268</td>
</tr>
<tr>
<td></td>
<td>lndi</td>
<td>0.53641</td>
</tr>
<tr>
<td></td>
<td>lpimin</td>
<td>0.29039</td>
</tr>
<tr>
<td></td>
<td>CONST</td>
<td>2.99512</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Coefficient</th>
<th><em>t</em> statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MATLAB</td>
<td>R</td>
</tr>
<tr>
<td>Fixed</td>
<td>lpcap</td>
<td>-0.04041</td>
</tr>
<tr>
<td></td>
<td>lpc</td>
<td>0.21904</td>
</tr>
<tr>
<td></td>
<td>lemp</td>
<td>0.66833</td>
</tr>
<tr>
<td></td>
<td>unemp</td>
<td>-0.00473</td>
</tr>
<tr>
<td></td>
<td>W*lgsp</td>
<td>0.19166</td>
</tr>
<tr>
<td></td>
<td>CONST</td>
<td>1.93735</td>
</tr>
<tr>
<td>Between</td>
<td>lpcap</td>
<td>0.17131</td>
</tr>
<tr>
<td></td>
<td>lpc</td>
<td>0.30163</td>
</tr>
<tr>
<td></td>
<td>lemp</td>
<td>0.58559</td>
</tr>
<tr>
<td></td>
<td>unemp</td>
<td>-0.00242</td>
</tr>
<tr>
<td></td>
<td>W*lgsp</td>
<td>-0.01082</td>
</tr>
<tr>
<td></td>
<td>CONST</td>
<td>1.70896</td>
</tr>
<tr>
<td>Random</td>
<td>lpcap</td>
<td>0.02286</td>
</tr>
<tr>
<td></td>
<td>lpc</td>
<td>0.29376</td>
</tr>
<tr>
<td></td>
<td>lemp</td>
<td>0.70864</td>
</tr>
<tr>
<td></td>
<td>unemp</td>
<td>-0.00648</td>
</tr>
<tr>
<td></td>
<td>W*lgsp</td>
<td>0.03547</td>
</tr>
<tr>
<td></td>
<td>CONST</td>
<td>1.90996</td>
</tr>
</tbody>
</table>
7. Conclusions

The new Panel Data Toolbox covers a wide variety of panel data models in the organized environment provided by MATLAB. Estimation methods include fixed, between and random effects, as well as instrumental variables models and spatial models.

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 and instrumental panel data methods. This positions the new toolbox as a valid self-contained alternative for panel data econometrics in MATLAB.

Future improvements aim at adding new econometric methods, including unbalanced and rotating panels, dynamic panel data models, and additional tests.

Acknowledgments

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 Science and Innovation (AP2010-1401).

References


Inmaculada C. Álvarez, Javier Barbero, José L. Zofío


**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/](http://www.paneldatatoolbox.com/)