UNIVERSITÀ CATTOLICA del Sacro Cuore

DIPARTIMENTO DI SCIENZE ECONOMICHE E SOCIALI

panelSUR: Two-way EC Model Estimation of SUR System on Unbalanced Panel Data in R

Laura Barbieri
Silvia Platoni

Quaderno n. 162/giugno 2024

VP VITA E PENSIERO

Università Cattolica del Sacro Cuore

DIPARTIMENTO DI SCIENZE ECONOMICHE E SOCIALI

panelSUR: Two-way EC Model Estimation of SUR System on Unbalanced Panel Data in R

Laura Barbieri
Silvia Platoni

Quaderno n. 162/giugno 2024

VP VITA E PENSIERO

Laura Barbieri, Dipartimento di Scienze Economiche e Sociali,
Università Cattolica del Sacro Cuore, Piacenza.

Silvia Platoni, Dipartimento di Scienze Economiche e Sociali,
Università Cattolica del Sacro Cuore, Piacenza.

laura.barbieri@unicatt.it

silvia.platoni@unicatt.it

I quaderni possono essere richiesti a:
Dipartimento di Scienze Economiche e Sociali,
Università Cattolica del Sacro Cuore
Via Emilia Parmense 84 - 29122 Piacenza - Tel. 0523 599 342
http://dipartimenti.unicatt.it/dises

dises-pc@unicatt.it

www.vitaepensiero.it

ISBN 978-88-343-5842-9

panelSUR: Two-way EC Model Estimation of SUR Systems on Unbalanced Panel Data in R

June 7, 2024

Abstract

Many statistical analyses (e.g., in econometrics) are based on seemingly unrelated regression (SUR) model on unbalanced panel data. The **panelSUR** package provides the possibility to estimate this kind of models within the R programming environment. This package can be used for the Generalized Least Squares (GLS) estimation of SUR systems on unbalanced panel in the "one-way case", where only the individual-specific effects are considered as well as in the "two-way case", where both the individual-specific and the time-specific effects are taken into account. Furthermore, the **panelSUR** package allows to take into account the possibility of cross-equation restrictions.

Keywords: Unbalanced panels; ECM; SUR; Heteroskedasticity.
JEL classification: C13; C23; C33.

1. INTRODUCTION

Many econometric analyses concern system of equations whose disturbance terms are likely to be correlated, because of unconsidered factors that contemporaneously influence the disturbance term of each equation. In order to obtain efficient estimates of the system coefficients, this contemporaneous correlation has to be taken into account and the equations should be simultaneously estimated. This could be generally done with a generalized least squares (GLS) estimator accounting for the covariance structure of the residuals, also known as seemingly unrelated regression (SUR) procedure (Zellner, 1962). It should be noted that a simultaneous estimation of the system equations also takes the important advantage of allowing possible cross-equation restrictions—which economic theory very often suggests—to be taken into account.

In a panel data framework, the efficient estimation of the system coefficients is even more complicated since the unconsidered factors influencing the equation disturbance terms may include individual-specific and time-specific effects. In this context, the error components (EC) model is the most frequently used approach and Baltagi (1980) and Magnus (1982) extended the estimation procedure of the single-equation model for balanced panels to the case of SUR system. Moreover, since the occurrence of missing observations—not all cross-sectional units are observed during all time periods—is common in practice, and unbalanced panels are the norm rather than the exception in large-scale survey data, Biørn (2004) proposed a parsimonious technique to estimate one-way SUR systems on unbalanced panel data.

Unbalanced panels present several inferential challenges, significant in econometric and statistical analyses, compared to balanced panels. First, sample selection bias problems can arise since missing data may not be random (i.e., if certain types of individuals or time periods are more likely to have missing data, the remaining data may not be representative of the entire population), and this can lead to biased estimates. Second, missing data reduces the amount of information available for estimation, leading to 1) less precise estimates and possible inconsistent estimates and 2) reduced statistical power.

Researchers need to carefully specify their models to account for the unbalanced nature of the data, for example, by using techniques such as imputation or data augmentation to mitigate some of the issues by filling in the gaps based on observed patterns. Moreover, standard estimation approaches (FE, RE, GMM, etc.) need to be sophisticated to handle the unbalanced structure. In summary, unbalanced panels introduce complexities that require careful handling to avoid biased and inconsistent estimates.

In order to be able to consider not only the individual-specific effect, but also the time-specific effect, Platoni et al. (2012a) further extended this technique to the case of the two-way EC model for unbalanced panels. They also proposed an extension of the Quadratic Unbiased Estimator (QUE) procedure suggested for the single equation case by Wansbeek and Kapteyn (1989) to a system of equations.

The **panelSUR** package provides the capability to estimate systems of linear equations in the R programming environment (RCoreTeam, 2022) by means of these techniques. Currently, the one-way and the two-way error component procedures based on the seminal Biørn (2004)'s procedure are implemented, as well as the two-way quadratic unbiased estimation procedure proposed by Platoni et al. (2012a). Accordingly to these approaches, the key element of the techniques is arranging the data such that individuals are grouped according to the number of times they are observed. Furthermore, the **panelSUR** package provides the possibility to take into account cross-equation restrictions on the coefficients.

Although systems of linear equations can be estimated in R via the **systemfit** package (as well as with several other statistical and econometric software packages, e.g. SAS, EViews, Stata), the **panelSUR** package has the considerably advantage to handle SUR system in the (unbalanced) panel data case.[1]

The paper is organized as follows. Section 2 reviews the statistical background of estimating equation systems. The implementation of the statistical procedures in R is briefly explained, via a simulated data set, in Section 3, and conclusions are finally given in Section 4.

2. THEORETICAL BACKGROUND

An unbalanced panel is characterized by a total of N observations, with n individuals (indexed $i = 1, \ldots, n$) observed over T periods (indexed $t = 1, \ldots, T$). Let T_i denote the number of times the individual i is observed, and n_t the number of individuals observed in period t. Hence $\sum_i T_i = \sum_t n_t = N$. The regression model is:

$$y_{it} = \mathrm{x}_{it}^\top \beta + \mu_i + \nu_t + u_{it} = \mathrm{x}_{it}^\top \beta + \varepsilon_{it}, \tag{1}$$

where x_{it} is a $k \times 1$ vector of explanatory variables, β a $k \times 1$ vector of parameters, μ_i the individual-specific effect, ν_t the time-specific effect, u_{it} the remainder error term, and $\varepsilon_{it} = \mu_i + \nu_t + u_{it}$ the composite error term.

As in Wansbeek and Kapteyn (1989) and Platoni et al. (2012a), the data are ordered on the n individuals in T consecutive sets, one for each period. Let D_t be the $n_t \times n$ matrix obtained from the $n \times n$ identity matrix I_n by omitting the rows corresponding

[1]Note that the **systemfit** package, in the panel data case, is able to handle only a single equation that is estimated for all individuals.

to individuals not observed in period t; thus, it is possible to define the $N \times n$ matrix $\Delta_\mu \equiv (\mathrm{D}_1^\top, \dots, \mathrm{D}_T^\top)$ and the $N \times T$ matrix $\Delta_\nu \equiv \mathrm{diag}[\mathrm{D}_t \iota_n] = \mathrm{diag}[\iota_{n_t}]$, where ι_n and ι_{n_t} are respectively $n \times 1$ and $n_t \times 1$ vectors of ones. Hence, using matrix notation, we can write:

$$\mathrm{y} = \mathrm{X}\beta + \Delta_\mu \mu + \Delta_\nu \nu + \mathrm{u} = \mathrm{X}\beta + \varepsilon, \tag{2}$$

where X is a $N \times k$ matrix of explanatory variables, μ a $n \times 1$ vector of individual-specific effects, ν a $T \times 1$ vector of time-specific effects, and u a $N \times 1$ vector of remainder error terms.

As in Biørn (2004) and Platoni et al. (2012a), individuals are grouped according to the number of times they are observed. Let n_p denote the number of individuals observed exactly in p periods, with $p = 1, \dots, T$. Hence $\sum_p n_p = n$ and $\sum_p (n_p p) = N$. The T groups of individuals are assumed to be ordered such that the n_1 individuals observed once come first, the n_2 individuals observed twice come second, etc. Hence with $C_p = \sum_{\tau=1}^{p} n_\tau$ being the cumulated number of individuals observed at most p times, the index sets of the individuals observed exactly p times can be written as $I_p = \{C_{p-1} + 1, \dots, C_p\}$.[2]

If k_m is the number of regressors for equation m, the total number of regressors for the system is $K = \sum_{m=1}^{M} k_m$. Stacking the M equations, indexed by $m = 1, \dots, M$, for the observation (i, t) we have:

$$\mathrm{y}_{it} = \mathrm{X}_{it}\beta + \mu_i + \nu_t + \mathrm{u}_{it} = \mathrm{X}_{it}\beta + \varepsilon, \tag{3}$$

where $\mathrm{X}_{it} = \mathrm{diag}[\mathrm{x}_{1it}, \dots, \mathrm{x}_{Mit}]$ is a $M \times K$ matrix of explanatory variables and $\beta = (\beta_1^\top, \dots, \beta_M^\top)^\top$ a $K \times 1$ vector of parameters.[3] The expected values of the $M \times 1$ vectors μ_i, ν_t, and u_{it} are

[2] Note that I_1 may be considered as a pure cross section and I_p, with $p \geq 2$, as a pseudo-balanced panel with p observations for each individual.

[3] As Biørn (2004) suggests, if the coefficient vectors are not disjoint across equations, we can redefine β as the complete $K \times 1$ coefficient vector (without

assumed to be zero and their covariance matrices be equal to Σ_μ, Σ_ν, and Σ_u. It follows that $\mathrm{E}(\varepsilon_{it}\varepsilon_{hs}^\top) = \delta_{ih}\Sigma_\mu + \delta_{ts}\Sigma_\nu + \delta_{ih}\delta_{ts}\Sigma_u$, with $\delta_{ih} = 1$ for $i = h$ and $\delta_{ih} = 0$ for $i \neq h$, $\delta_{ts} = 1$ for $t = s$, and $\delta_{ts} = 0$ for $t \neq s$ (see Appendix A for more details).

Referring to individual $i \in I_p$, the $pM \times 1$ vector of independent variables $\mathrm{y}_{i(p)} \equiv (\mathrm{y}_{i1}^\top, \ldots, \mathrm{y}_{ip}^\top)^\top$, the $pM \times K$ matrix of explanatory variables $\mathrm{X}_{i(p)} \equiv (\mathrm{X}_{i1}^\top, \ldots, \mathrm{X}_{ip}^\top)^\top$, and the $pM \times 1$ vector of composite error terms $\varepsilon_{i(p)} \equiv (\varepsilon_{i1}^\top, \ldots, \varepsilon_{ip}^\top)^\top$ have to be considered. Therefore in Platoni et al. (2012a) the model is written as:

$$\mathrm{y}_{i(p)} = \mathrm{X}_{i(p)}\beta + (\iota_p \otimes \mu_i) + \nu_{i(p)} + \mathrm{u}_{i(p)} = \mathrm{X}_{i(p)}\beta + \varepsilon_{i(p)}, \quad (4)$$

where ι_p is a $p \times 1$ vector of ones and $\nu_{i(p)} \equiv \Delta_{i(p)}\nu$ a $pM \times 1$ vector of time-specific errors for the individual $i \in I_p$, with $\Delta_{i(p)}$ a $pM \times TM$ matrix which detects in which period t the individual i of the group p is observed.

The $pM \times pM$ variance-covariance matrix of the composite error terms $\varepsilon_{i(p)}$ is given by:

$$\Omega_p = \mathrm{E}_p \otimes (\Sigma_u + \Sigma_\nu) + \bar{\mathrm{J}}_p \otimes (\Sigma_u + \Sigma_\nu + p\Sigma_\mu), \quad (5)$$

where $\mathrm{E}_p = \mathrm{I}_p - \bar{\mathrm{J}}_p$, I_p is an identity matrix of dimension p, $\bar{\mathrm{J}}_p = \frac{1}{p}\mathrm{J}_p$, and J_p is a matrix of ones of dimension p. Since the $p \times p$ matrices E_p and $\bar{\mathrm{J}}_p$ are symmetric, idempotent, and have orthogonal columns, the inverse of the $pM \times pM$ variance-covariance matrix is:

$$\Omega_p^{-1} = \mathrm{E}_p \otimes (\Sigma_u + \Sigma_\nu)^{-1} + \bar{\mathrm{J}}_p \otimes (\Sigma_u + \Sigma_\nu + p\Sigma_\mu)^{-1}. \quad (6)$$

---

duplication) and the $M \times K$ regression matrix as $\mathrm{X}_{it} = (\mathrm{x}_{1it}^\top, \mathrm{x}_{2it}^\top, \ldots, \mathrm{x}_{Mit}^\top)^\top$ where the k^{th} element of the $1 \times k_m$ vector x_{mit} (i) contains the observation on the variable in the m^{th} equation which corresponds to the k^{th} coefficient in β or (ii) is zero if the k^{th} coefficient does not occur in the m^{th} equation.

If we assume that Σ_u, Σ_μ, and Σ_ν are known, we can write the GLS estimator for β as the problem of minimizing:

$$\sum_{p=1}^{T}\sum_{i\in I_p}\varepsilon_{i(p)}^{\top}\Omega_p^{-1}\varepsilon_{i(p)} \tag{7}$$

If we apply GLS on the observations for the individuals observed p times we obtain:

$$\hat{\beta}_p^{GLS} = \left(\sum_{i\in I_p}\mathrm{X}_{i(p)}^{\top}\Omega_p^{-1}\mathrm{X}_{i(p)}\right)^{-1}\left(\sum_{i\in I_p}\mathrm{X}_{i(p)}^{\top}\Omega_p^{-1}\mathrm{y}_{i(p)}\right), \tag{8}$$

while the full GLS estimator is:

$$\hat{\beta}^{GLS} = \left(\sum_{p=1}^{T}\sum_{i\in I_p}\mathrm{X}_{i(p)}^{\top}\Omega_p^{-1}\mathrm{X}_{i(p)}\right)^{-1}\left(\sum_{p=1}^{T}\sum_{i\in I_p}\mathrm{X}_{i(p)}^{\top}\Omega_p^{-1}\mathrm{y}_{i(p)}\right). \tag{9}$$

Platoni et al. (2012a) propose two procedures to estimate the three error component variance-covariance matrices of the two-way SUR system Σ_u, Σ_μ, and Σ_ν: (i) the first one is achieved by modifying the QUE procedure suggested by Wansbeek and Kapteyn (1989) for the single equation case, (ii) the second by modifying the within-between (hereinafter, WB) procedure suggested by Biørn (2004) for the one-way SUR system.

2.1. The QUE procedure

The QUE procedure considers the $N \times 1$ Fixed Effect (FE) residuals $\mathrm{e}_m \equiv \mathrm{y}_m - \mathrm{X}_m\hat{\beta}_m^W$ from the Within (W) estimator:

$$\hat{\beta}_m^W = \left(\mathrm{X}_m^{\top}\mathrm{Q}_\Delta\mathrm{X}_m\right)^{-1}\mathrm{X}_m^{\top}\mathrm{Q}_\Delta\mathrm{y}_m \tag{10}$$

for the equation $m = 1, \ldots, M$, where X_m does not include the intercept, and thus it is a matrix of dimension $N \times (k_m - 1)$. The

$N \times N$ matrix Q_Δ on which the two-way EC model transformation is based is:

$$\mathrm{Q}_\Delta = \mathrm{Q}_A - \mathrm{P}_B = \mathrm{Q}_A - \mathrm{Q}_A \Delta_\nu \mathrm{Q}^- \Delta_\nu^\mathrm{T} \mathrm{Q}_A, \tag{11}$$

with $\mathrm{Q}_A = \mathrm{I}_N - \mathrm{P}_A$, $\mathrm{P}_A = \Delta_\mu \Delta_n^{-1} \Delta_\mu^\mathrm{T}$, the $n \times n$ matrix $\Delta_n \equiv \Delta_\mu^\top \Delta_\mu$, the $T \times T$ matrix $\mathrm{Q} = \Delta_\nu^\mathrm{T} \mathrm{Q}_A \Delta_\nu$ whose Q^- is the generalized inverse (see Wansbeek and Kapteyn, 1989; Davis, 2002). However, if we assume that the $N \times k_m$ matrix X_m in (10) contains a vector of ones, the $N \times 1$ centered residuals $\mathrm{f}_m \equiv \mathrm{E}_N \mathrm{e}_m = \mathrm{e}_m - \bar{e}_m$ have to be considered, where $\bar{e}_m = \frac{1}{N} \sum_{i=1}^{n} \sum_{t=1}^{T_i} e_{mit} = \frac{1}{N} \sum_{t=1}^{T} \sum_{i=1}^{n_t} e_{mit}$, and $\mathrm{E}_N = \mathrm{I}_N - \bar{\mathrm{J}}_N$, with I_N being an identity matrix of dimension N, $\bar{\mathrm{J}}_N = \frac{1}{N} \mathrm{J}_N$, and J_N a matrix of ones of dimension N[4].

As in Platoni et al. (2012a), the adapted QUEs for $\sigma_{u_{mj}}^2$, $\sigma_{\mu_{mj}}^2$, and $\sigma_{\nu_{mj}}^2$ is obtained by equating:

$$q_{N_{mj}} \equiv \mathrm{f}_j^\top \mathrm{Q}_\Delta \mathrm{f}_m, \tag{12a}$$
$$q_{n_{mj}} \equiv \mathrm{f}_j^\top \Delta_\mu \Delta_n^{-1} \Delta_\mu^\top \mathrm{f}_m, \tag{12b}$$
$$q_{T_{mj}} \equiv \mathrm{f}_j^\top \Delta_\nu \Delta_T^{-1} \Delta_\nu^\top \mathrm{f}_m, \tag{12c}$$

with the $T \times T$ matrix $\Delta_T \equiv \Delta_\nu^\top \Delta_\nu$, to their expected values:

$$\mathrm{E}\left(q_{N_{mj}}\right) = \left(N - T - n + 1 + k_{mj} - k_m - k_j\right) \cdot \hat{\sigma}_{u_{mj}}^2, \tag{13a}$$
$$\mathrm{E}\left(q_{n_{mj}}\right) = \left(n + k_{n_{mj}} - k_{0_{mj}} - 1\right) \cdot \hat{\sigma}_{u_{mj}}^2 + \left(N - \lambda_\mu\right) \cdot \hat{\sigma}_{\mu_{mj}}^2 + \left(n - \lambda_\nu\right) \cdot \hat{\sigma}_{\nu_{mj}}^2, \tag{13b}$$
$$\mathrm{E}\left(q_{T_{mj}}\right) = \left(T + k_{T_{mj}} - k_{0_{mj}} - 1\right) \cdot \hat{\sigma}_{u_{mj}}^2 + \left(T - \lambda_\mu\right) \cdot \hat{\sigma}_{\mu_{mj}}^2 + \left(N - \lambda_\nu\right) \cdot \hat{\sigma}_{\nu_{mj}}^2, \tag{13c}$$

[4]When there is an intercept in the regression, the residuals e_m do not have a zero mean, so we need to refer to the centered residuals $\mathrm{f}_m \equiv \mathrm{E}_N \mathrm{e}_m$ rather than e_m (see Wansbeek and Kapteyn, 1989).

with $k_{mj} \equiv \mathrm{tr}((\mathrm{X}_m^\top \mathrm{Q}_\Delta \mathrm{X}_m)^{-1}\mathrm{X}_m^\top \mathrm{Q}_\Delta \mathrm{X}_j(\mathrm{X}_j^\top \mathrm{Q}_\Delta \mathrm{X}_j)^{-1}\mathrm{X}_j^\top \mathrm{Q}_\Delta \mathrm{X}_m)$, $k_{0_{mj}} \equiv \frac{1}{N}\iota_N^\top \mathrm{X}_m(\mathrm{X}_m^\top \mathrm{Q}_\Delta \mathrm{X}_m)^{-1}\mathrm{X}_m^\top \mathrm{Q}_\Delta \mathrm{X}_j(\mathrm{X}_j^\top \mathrm{Q}_\Delta \mathrm{X}_j)^{-1}\mathrm{X}_j^\top \iota_N$, $k_n \equiv \mathrm{tr}[(\mathrm{X}^\top \mathrm{Q}_\Delta \mathrm{X})^{-1}\mathrm{X}^\top \Delta_\mu \Delta_n \Delta_\mu^\top \mathrm{X}]$, $\lambda_\mu \equiv \frac{1}{N}\iota_N^\top \Delta_\mu \Delta_\mu^\top \iota_N = \frac{1}{N}\sum_{i=1}^n T_i^2$, $k_T \equiv \mathrm{tr}[(\mathrm{X}^\top \mathrm{Q}_\Delta \mathrm{X})^{-1}\mathrm{X}^\top \Delta_\nu \Delta_T \Delta_\nu^\top \mathrm{X}]$, and $\lambda_\nu \equiv \frac{1}{N}\iota_N^\top \Delta_\nu \Delta_\nu^\top \iota_N = \frac{1}{N}\sum_{t=1}^T n_t^2$.

2.2. The WB procedure

As the QUE procedure, the WB procedure considers the consistent $M \times 1$ FE residuals $\mathrm{e}_{it} \equiv \mathrm{y}_{it} - \mathrm{X}_{it}\hat{\beta}^W$ for the individual i in period t from the W estimator:

$$\hat{\beta}^W = \left(\mathrm{X}_{it}^\top \mathrm{Q}_\Delta \mathrm{X}_{it}\right)^{-1} \mathrm{X}_{it}^\top \mathrm{Q}_\Delta \mathrm{y}_{it}, \tag{14}$$

where X_{it} is a matrix of dimension $M \times (K - M)$.[5] As above, if we assume that the $M \times K$ matrix X_{it} in (14) contains M vectors of ones (a vector of ones for each equation m), then we have to define the $M \times 1$ consistent centered residuals $\mathrm{f}_{it} = \mathrm{e}_{it} - \bar{\mathrm{e}}$, where $\bar{\mathrm{e}} \equiv (\bar{e}_1, \ldots, \bar{e}_M)^\top$.

Therefore, the $M \times M$ matrices of within individuals, between individuals, and between times (co)variations in the f's of the

[5]To obtain consistent estimates of the variance-covariance matrices Σ_u, Σ_μ, and Σ_ν, we need consistent residuals (see Biørn, 2004), and thus for coherence in the within-between procedure we use the same FE residuals on which the QUE procedure is based.

different M equations are the following:

$$\mathrm{W}_f = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\left(\mathrm{f}_{it} - \bar{\mathrm{f}}_{i\cdot} - \bar{\mathrm{f}}_{\cdot t}\right)\left(\mathrm{f}_{it} - \bar{\mathrm{f}}_{i\cdot} - \bar{\mathrm{f}}_{\cdot t}\right)^\top, \tag{15a}$$

$$\mathrm{B}_f^C = \sum_{i=1}^{n} T_i\left(\bar{\mathrm{f}}_{i\cdot} - \bar{\mathrm{f}}\right)\left(\bar{\mathrm{f}}_{i\cdot} - \bar{\mathrm{f}}\right)^\top, \tag{15b}$$

$$\mathrm{B}_f^T = \sum_{t=1}^{T} n_t\left(\bar{\mathrm{f}}_{\cdot t} - \bar{\mathrm{f}}\right)\left(\bar{\mathrm{f}}_{\cdot t} - \bar{\mathrm{f}}\right)^\top, \tag{15c}$$

where for each equation m we have $\bar{f}_{mi\cdot} = \frac{1}{T_i}\sum_{t=1}^{T_i} f_{mit}$, $\bar{f}_{m\cdot t} = \frac{1}{n_t}\sum_{i=1}^{n_t} f_{mit}$ and $\bar{f}_m = \frac{1}{N}\sum_{i=1}^{n}\sum_{t=1}^{T_i} f_{mit} = \frac{1}{N}\sum_{i=1}^{n}(T_i\bar{f}_{mi\cdot})$ or $\bar{f}_m = \frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{n_t} f_{mit} = \frac{1}{N}\sum_{t=1}^{T}(n_t\bar{f}_{m\cdot t})$. As in Platoni et al. (2012a), the between times (co)variation B_f^T is needed to adapt the Biørn (2004)'s procedure to the two-way EC model. Therefore, consistent and unbiased estimators of Σ_u, Σ_μ, and Σ_ν can be obtained from:

$$\mathrm{E}\left(\mathrm{W}_f\right) = \left(N - n - T\right)\hat{\Sigma}_u, \tag{16a}$$

$$\mathrm{E}\left(\mathrm{B}_f^C\right) = \left(N - \frac{1}{N}\sum_{i=1}^{n} T_i^2\right)\hat{\Sigma}_\mu + \left(n-1\right)\hat{\Sigma}_u, \tag{16b}$$

$$\mathrm{E}\left(\mathrm{B}_f^T\right) = \left(N - \frac{1}{N}\sum_{t=1}^{T} n_t^2\right)\hat{\Sigma}_\nu + \left(T-1\right)\hat{\Sigma}_u. \tag{16c}$$

2.3. Goodness of fit

In order to measure the goodness of fit of each single equation of the system we refer in a first attempt to the traditional multiple R^2 values

$$R_m^2 = 1 - \frac{\hat{\varepsilon}_m^\top\hat{\varepsilon}_m}{(\mathrm{y}_m - \bar{\mathrm{y}}_m)^\top(\mathrm{y}_m - \bar{\mathrm{y}}_m)} \tag{17}$$

where R_m^2 is the R^2 value of the mth equation, $\bar{y}_m$ is the mean value of y_m, and $\hat{\varepsilon}_m$ is obtained as the difference between y_m and its estimate obtained as $\hat{y}_m = X_{it}\hat{\beta}$. This R^2, also known as "R^2 overall", could be viewed as the squared correlation between y_m and $\hat{y}_m$ and being equal to the fraction of the variation in y_m explained by the estimated equation.

3. PACKAGE DESCRIPTION AND ILLUSTRATIVE EXAMPLES

3.1. Source code

The source code of the **panelSUR** package is publicly available for download from the Comprehensive R Archive Network (CRAN, `http://CRAN.R-project.org/`). The required packages are **MASS**, **plm**, **matlib**, **fastmatrix**, and **formula.tools**.

The basic functionality of **panelSUR** package is provided by the function `SURest`. Additionally, the package includes several internal helper functions:

- `prepareData` to prepare the data for use by elaborating all the necessary information;
- `preliminaryEstimate` to obtain the preliminary FE estimates;
- `obtainSigmas` and `system` respectively to compute the Σ matrices presented in previous Section 2 and to solve the system.

If `SURest` is applied, it calls all the internal helper functions to obtain the final results.

3.2. The basic function SURest

The `SURest` function is the core of the **panelSUR** package. It finally allows to implement the three different estimation techniques presented above to systems of linear equations. The user interface and the returned object of this function are very similar to those of other R functions (i.e. `lm` and `systemfit`) and make its usage as easy as possible for R users. The econometric estimation is done by applying the formulas in Sections 1 and 2, and if restrictions on the coefficients are specified symbolically, a restricted matrix is automatically generated into the program.

The `SURest` function returns a list of objects that belong to the entire system of equations. Moreover, a `printSUR()` function is provided in order to display the main elements of the estimation results in a summary table.

3.3. Using SURest

For a simple illustration of the package, the `SURdata` dataset included in the **panelSUR** package is firstly used.[6]. Dataset consists of a simulated unbalanced panel comprising 100 individuals observed across four time periods for a total of 220 observations ($n = 100$, $T = 4$, $N = 220$). For each group p we have the following number of individuals: $n_1 = 34$, $n_2 = 28$, $n_3 = 22$, $n_4 = 16$ (and thus $N = 220$). Data on three independent variables (Y_i, $i = 1, 2, 3$) and three dependent variables (X_i, $i = 1, 2, 3$) are generated.

As referring model, the following two equations system ($M = 2$) is considered:

$$\begin{cases} y_1 = \beta_{10} + \beta_{11} x_1 + \beta_{12} x_2 + \varepsilon_1 \\ y_2 = \beta_{20} + \beta_{21} x_1 + \beta_{22} x_2 + \beta_{23} x_3 + \varepsilon_2 \end{cases} \tag{18}$$

To begin the analysis, data are loaded by:

[6]This dataset is a shortened version of the simulated datasets already analyzed in Platoni et al. (2012a) and that will be extensively presented in Section 4.

```
R> library("panelSUR")
R> data("SURdata", package = "panelSUR")
R> head(data)

# A tibble: 6 x 9
    Obs IND     TIME     Y1     Y2    Y3    X1    X2    X3
  <dbl> <chr>  <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
1     1 COD001  3001 -32.6   11.5   21.3 2.79  5.23  0.845
2     2 COD002  3000   6.49  71.6   19.1 0.303 2.99  1.53
3     3 COD003  3003  37.3   18.1   42.1 3.51  0.823 3.32
4     4 COD004  3002  69.4   71.2   31.0 1.11  2.82  2.55
5     5 COD005  3001 -13.2  -24.5   32.3 4.46  1.36  3.82
6     6 COD006  3000  93.3    8.88 -16.1 4.76  2.20  3.24
```

and the two equations of system (18) are specified as object of the class `formula()` and put in a `list()` as follows:

```
R> eq1 <- Y1~X1+X2
R> eq2 <- Y2~X1+X2+X3
R> eqlist<-list(eq1,eq2)
```

Note that the constant term must be always included in each equation. If there are any, we can also specify our coefficient constraints (see Section 4).

Finally, the data must be provided in a transformed data frame of class `pdata.frame`[7] which can be created with the function `pdata.frame` from the R package **plm** (Croissant and Millo, 2008). The resulting `pdata.frame` is sorted by the individual index, then by the time index:

```
R> library(plm)
R> datap <- pdata.frame(SURdata, index=c("IND", "TIME"))
```

Having defined all needed arguments (see Table 1), the one-way WB estimation of system (19)—that here we named `mod0` for simplicity—can be obtained via the command `SURest` as follows:

[7]Note that the data must be in "long format" (different individuals arranged below each other), not in the "wide format" (different individuals arranged next to each other).

```
R> mod0 <- SURest(eqlist=eqlist, data=datap, method="1wayWB")
```

Arguments	Description
`eqlist`	is the `list()` of the equations that make up the SUR system.
`restrictions`	is the vector containing the constraints on the equation coefficients.
`method`	the estimation method to be used, one of "1wayWB", "2wayWB", or "2wayQUE".
`data`	a data frame of class 'pdata.frame' (mandatory).

Table 1
List of arguments for the function `SURest()`.

The results of the `SURest` function are stored in a `list()` object and can be simply extracted as usual. These results include the vector of the coefficient estimates of the system equations (`Estimate`) and their main attributes (`std_error`, `tstat`, `pvalue`, etc.). Nevertheless, in order to easily display the estimation result in a summary table, a `printSUR()` function is provided.

In particular, by applying the `printSUR()` command to the model estimated above, we can easily display the results of the one-way WB estimations procedure as follows

```
R> printSUR(mod0)

SUR estimation results
Method: One-way WB

Unbalanced Panel: n = 100, T = 1-4, N = 220

=============  ==============  ============  ===========  ========
 Coefficient      Estimate       Std.Error     t-value     p-value
=============  ==============  ============  ===========  ========
const               19.41718       4.75701      4.08180   0.00006
X1                   4.65553       1.47023      3.16653   0.00176
X2                  -3.81046       1.42422     -2.67547   0.00804
const               10.46694       4.09608      2.55535   0.01130
X1                  -5.49576       1.27658     -4.30506   0.00002
X2                   7.46762       1.27890      5.83910   0.00000
```

```
X3                     -0.40308       1.39981     -0.28796   0.77366
=============  ==============  ============  ===========  ========

Multiple R-Squared for single equation
R1 = 0.981545, R2 = 0.844404
```

The output reports the main information about the estimated model, such as the method used, the dataset dimensions, the results obtained in terms of estimated coefficient, and the overall R^2 values for each equation included in the `eqlist`.

Some interesting object that can be extracted from the `SUREst` object includes the error variance-covariance matrices. In particular, we can see the remainder error variance-covariance matrix Σ_u obtained implementing equation (16a)

```
mod0$Sigma_u
         [,1]     [,2]
[1,] 284.9563 117.5918
[2,] 117.5918 193.8878
```

the individual error variance-covariance matrix Σ_μ (equation (16b))

```
mod0$Sigma_mu
           [,1]      [,2]
[1,] 1054.96305 -24.10922
[2,]  -24.10922 757.32873
```

and, finally, the time error variance-covariance matrix Σ_ν (equation (16c))

```
mod0$Sigma_nu
     [,1] [,2]
[1,]    0    0
[2,]    0    0
```

Note that, in this case, being a one-way estimation, the Σ_ν matrix is obviously null.

The program also provides the possibility to obtain the two-way WB and the two-way QUE estimations by Platoni et al. (2012a) (see below).

3.4. A replication example

In this section, we reproduce some results presented in Platoni et al. (2012a). To this aim we use the last one of the 150 simulated datasets already analyzed in that paper and consisting of an unbalanced panel with a large number of individuals ($n = 4000$) extended over a rather long time period ($T = 8$).[8] In order to construct this unbalanced panel, the procedure currently used for rotating panels, in which there is approximately the same number of individuals every year, has been used: a fixed percentage of individuals (20% in this case[9]) is replaced each year, but they can re-enter the sample in the following years. Thus, for each group p we have the following number of individuals: $n_1 = 962$, $n_2 = 769$, $n_3 = 615$, $n_4 = 492$, $n_5 = 394$, $n_6 = 315$, $n_7 = 252$, and $n_8 = 201$ (and thus $N = 13545$).

As referring model, the same three equations system ($M = 3$) considered in Platoni et al. (2012a) is estimated:

$$\begin{cases} y_1 &= \beta_{10} &+\beta_{11}x_1 &+\beta_{12}x_2 & &+\varepsilon_1 \\ y_2 &= \beta_{20} &+\beta_{21}x_1 &+\beta_{22}x_2 &+\beta_{23}x_3 &+\varepsilon_2 \\ y_3 &= \beta_{30} & &+\beta_{32}x_2 &+\beta_{33}x_3 &+\varepsilon_3 \end{cases} \tag{19}$$

where the true values of the coefficient vectors are:

$$\begin{cases} \boldsymbol{\beta}_1 = (15, 6, -3, 0)^\top \\ \boldsymbol{\beta}_2 = (10, -3, 8, -2)^\top \\ \boldsymbol{\beta}_3 = (20, 0, -2, 5)^\top. \end{cases} \tag{20}$$

Then the following cross-equation restrictions are allowed:

$$\begin{aligned} \beta_{12} &= \beta_{21} \\ \beta_{23} &= \beta_{32} \end{aligned} \tag{21}$$

[8] The dataset is available from the authors upon request.

[9] Also in Wansbeek and Kapteyn (1989) each period 20% of the households in the panel is removed randomly.

Moreover, the following variance-covariance matrices are considered[10]:

$$\Sigma_u = \begin{bmatrix} 86.28 & 17.39 & \text{-}5.94 \\ & 77.98 & 7.53 \\ & & 56.46 \end{bmatrix}, \; \Sigma_\mu = \begin{bmatrix} 968.5 & \text{-}88.2 & 21.5 \\ & 725.2 & \text{-}55.0 \\ & & 513.4 \end{bmatrix}, \\ \text{and } \Sigma_\nu = \begin{bmatrix} 87.52 & 15.81 & \text{-}4.65 \\ & 79.97 & 5.89 \\ & & 53.22 \end{bmatrix}. \tag{22}$$

Finally, the independent variables' values x_{kit} ($k = 1, 2, 3$) are generated according to a modified version of the Data Generating Process (DGP) introduced by Nerlove (1971) and used, among others, by Baltagi (1981), Wansbeek and Kapteyn (1989), and Platoni et al. (2012a):

$$x_{kit} = 0.1t + 0.5x_{kit-1} + \omega_{kit}, \quad k = 1, 2, 3,$$

with ω_{kit} following the uniform distribution $[-\frac{1}{2}, \frac{1}{2}]$ and $x_{ki0} = 5 + 10\omega_{ki0}$.

To begin the analysis, data are loaded by:

```
R> library("panelSUR")
R> load("Platoni_SURdata.RData")
```

and the three equations of system (19) are specified as follows:

```
R> eq1 <- Y1~X1+X2
R> eq2 <- Y2~X1+X2+X3
R> eq3 <- Y3~X2+X3
R> eqlist<-list(eq1,eq2,eq3)
```

In order to specify the vector of the coefficient restrictions (21)[11], each constraint should be put into quotation marks and

[10] The three variance-covariance matrices, used in Platoni et al. (2012a), have been randomly generated using the `sprandsym` command in MatLab, that produces positive-definite symmetric matrices with all non-zero entries.

[11] This version of the package allows considering only simple restrictions involving equality between two parameters, and not linear combinations involving more than two parameters.

each coefficient should be indicated as equation_name$variable_name. Additionally, any spaces should be excluded from the restrictions definition. If one of the constraints includes one of the intercept terms, the variable_name will be simply const. In our example, the coefficient constraints (21) should be written as:

```
R> constraints <- c("eq1$X2=eq2$X1","eq2$X3=eq3$X2")
```

Finally, data should be sorted by the individual index, then by the time index:

```
R> datap <- pdata.frame(Platoni_SURdata, index=c("IND", "TIME"))
```

Having defined all needed arguments, the QUE estimation of system (19)—that here we named `mod1` for simplicity—can be obtained via the command `SURest` as follows:

```
R> mod1 <- SURest(eqlist=eqlist, restrictions=constraints,
                  data=datap, method="2wayQUE")
R> printSUR(mod1)

SUR estimation results
Method: Two-way QUE

Unbalanced Panel: n = 4000, T = 1-8, N = 13545

=============  ==============  ============  ===========  ========
 Coefficient      Estimate       Std.Error     t-value    p-value
=============  ==============  ============  ===========  ========
const                12.30835       0.54546     22.56491   0.00000
X1                    6.24090       0.11898     52.45530   0.00000
X2                   -3.06568       0.09255    -33.12306   0.00000
const                14.51394       0.49572     29.27834   0.00000
X2                    8.08518       0.13572     59.57188   0.00000
X3                   -1.94606       0.09834    -19.78990   0.00000
const                17.36384       0.42779     40.58982   0.00000
X3                    5.00973       0.12447     40.24910   0.00000
=============  ==============  ============  ===========  ========

Multiple R-Squared for single equation
R1 = 0.958486, R2 = 0.920978, R3 = 0.951423
```

Note that in the output list of Estimate there is no repetition of estimated coefficients. This means that the coefficients of variables subject to constraints are reported only referring to the variable appearing first in the system of equations.

The estimates are obtained by applying formula (10) after incorporating coefficient constraints into the regressor matrix. This is achieved by structuring the regressor matrix so that the variables associated by a constraint are arranged into the same column (see footnote 3).

It is possible to appreciate the validity of the results by comparing our estimated coefficients with their true values in (20). In addition, the obtained error variance-covariance matrices could be compared with the originating matrices in (22). Specifically, we have the following remainder error variance-covariance matrix Σ_u obtained implementing equation (13a)

```
mod1$Sigma_u
         [,1]      [,2]      [,3]
[1,] 88.13754 17.800357 -5.039290
[2,] 17.80036 77.878941  8.120942
[3,] -5.03929  8.120942 56.929589
```

the individual error variance-covariance matrix Σ_μ (equation (13b))

```
mod1$Sigma_mu
           [,1]       [,2]      [,3]
[1,]  949.27274 -109.42127  35.73197
[2,] -109.42127  709.57855 -41.85827
[3,]   35.73197  -41.85827 504.52449
```

and, finally, the time error variance-covariance matrix Σ_ν (equation (13c))

```
mod1$Sigma_nu
          [,1]      [,2]      [,3]
[1,]  10.76998 -10.11352  15.27600
[2,] -10.11352  34.81976 -25.70210
[3,]  15.27600 -25.70210  48.50222
```

As already demonstrated in Platoni et al. (2012a), the matrices obtained by this procedure are closed to the target. Clearly, some deviation occurs in our case with respect to their results. This misalignment is unsurprising, considering that we employ a single-run estimation approach rather than the full multi-run simulation described in their paper.

Finally, the estimates by the two-way WB procedure can be obtained

```
R> mod2 <- SURest(eqlist=eqlist, restrictions=constraints,
                  data=datap, method="2wayWB")
R> printSUR(mod2)

SUR estimation results
Method: Two-way WB

Unbalanced Panel: n = 4000, T = 1-8, N = 13545

=============  ==============  ============  ===========  ========
 Coefficient      Estimate       Std.Error     t-value    p-value
=============  ==============  ============  ===========  ========
const                12.30409       0.54704     22.49210   0.00000
X1                    6.24442       0.12144     51.42015   0.00000
X2                   -3.06697       0.09478    -32.35809   0.00000
const                14.52096       0.50050     29.01318   0.00000
X2                    8.08057       0.14228     56.79295   0.00000
X3                   -1.94414       0.10271    -18.92863   0.00000
const                17.35594       0.43366     40.02158   0.00000
X3                    5.01203       0.13013     38.51701   0.00000
=============  ==============  ============  ===========  ========

Multiple R-Squared for single equation
R1 = 0.958486, R2 = 0.920978, R3 = 0.951421
```

As for the two-way QUE, also for the two-way WB method from the `SURest` object it is possible to obtain the variance-covariance matrices:

```
mod2$Sigma_u
           [,1]       [,2]       [,3]
[1,] 91.230252 14.5567609 -1.1585143
[2,] 14.556761 87.2447380  0.9846959
[3,] -1.158514  0.9846959 66.9045431

mod2$Sigma_mu
           [,1]       [,2]      [,3]
[1,]  950.21658 -110.18184  37.18784
[2,] -110.18184  712.77779 -44.12748
[3,]   37.18784  -44.12748 509.85636

mod2$Sigma_nu
          [,1]      [,2]      [,3]
[1,]  11.03139 -10.14194  15.28361
[2,] -10.14194  35.01100 -25.70949
[3,]  15.28361 -25.70949  48.63623
```

Obviously, the estimated coefficients and variance-covariance matrices obtained by the two different methodologies are very close both to each other and to their true values.

3.5. An empirical application

Platoni et al. (2012b) estimate the following $M = 10$ system of output supply, input demand, and land allocation equations:

$$\begin{cases} q_j = & \delta_j + \sum_{k=1}^{m} \delta_{jk}^{p} \cdot \bar{p}_k + \sum_{k=1}^{\ell-1} \delta_{jk}^{w} \cdot \bar{w}_k + \sum_{k=1}^{m_p} \delta_{jk}^{b} \cdot \bar{b}_k \\ & + \sum_{k=1}^{K} \delta_{jk}^{V} \cdot \bar{V}_k + \sum_{k=1}^{L} \delta_{jk}^{v} \cdot \bar{v}_k, \quad j = 1, \ldots, m, \\ g_h = & -\gamma_h - \sum_{k=1}^{m} \gamma_{hk}^{p} \cdot \bar{p}_k - \sum_{k=1}^{\ell-1} \gamma_{hk}^{w} \cdot \bar{w}_k - \sum_{k=1}^{m_p} \gamma_{hk}^{b} \cdot \bar{b}_k \\ & - \sum_{k=1}^{K} \gamma_{hk}^{V} \cdot \bar{V}_k - \sum_{k=1}^{L} \gamma_{hk}^{v} \cdot \bar{v}_k, \quad h = 1, \ldots, l - \ell, \\ s_r = & \lambda_r + \sum_{k=1}^{m} \lambda_{rk}^{p} \cdot \bar{p}_k + \sum_{k=1}^{\ell-1} \lambda_{rk}^{w} \cdot \bar{w}_k + \sum_{k=1}^{m_p} \lambda_{rk}^{b} \cdot \bar{b}_k \\ & + \sum_{k=1}^{K} \lambda_{rk}^{V} \cdot \bar{V}_k + \sum_{k=1}^{L} \lambda_{rk}^{v} \cdot \bar{v}_k, \quad r = 1, \ldots, m_p, \end{cases} \qquad (23)$$

where $m = 5$ is the number of crops (durum wheat, maize, other cereals, oilseeds, and other arable crops), $\ell = 1$ is the number of variable inputs (crop inputs), and $m_p = 4$ is the number of crops included in the arable crop regime (durum wheat, maize, other cereals, and oilseeds). Moreover, the standard symmetry and reciprocity properties are imposed with the following cross-equation parametric restrictions: $\sum_{j=1}^{m-1} j = 10$ restrictions $\delta^p_{jk} = \delta^p_{kj}$ related to $m = 5$ outputs, $\sum_{h=1}^{\ell-1} h = 0$ restrictions $\gamma^w_{hk} = \gamma^w_{kh}$ related to $\ell = 1$ input, and $\sum_{r=1}^{m_p-1} r = 6$ restrictions $\lambda^b_{rk} = \lambda^b_{kr}$ related to $m_p = 4$ land allocations:

$$\begin{array}{lll} \delta^p_{jk} = \delta^p_{kj}, & j = 1, \ldots, m-1 = 4, & k = j+1 \\ \lambda^b_{rk} = \lambda^b_{kr}, & r = 1, \ldots, m_p - 1 = 3, & k = r+1 \end{array} \tag{24}$$

The data used for the analysis in Platoni et al. (2012b) are taken from the EU FADN database for the period 1994-2003 and refer to a sample of specialized Italian arable crop farms: the database is an unbalanced panel of $n = 14288$ farms observed over $T = 10$ years, for a total of $N = 34140$ observations. Furthermore, for each group p the number of farms is: $n_1 = 14288$, $n_2 = 7526$, $n_3 = 4869$, $n_4 = 3277$, $n_5 = 1826$, $n_6 = 1101$, $n_7 = 685$, $n_8 = 382$, $n_9 = 160$, and $n_{10} = 26$.[12]

All the results achieved by the authors using a TSP code specifically tailored for the case under study can be reproduced using the **panelSUR** package. In fact, as an illustrative example, below we replicate the estimation of the fourth equation (i.e., the oilseeds) in both the two-way WB and two-way QUE cases. The replicated results match exactly.

[12]Note also that Platoni et al. (2012b) solve the issue of heteroscedasticity of the remainder error term due to censoring (see, Shonkwiler and Yen, 1999) by adapting to the two-way case the error variance formulated by Tauchmann (2005) for the one-way case.

```
R> mod5 <- SURest(eqlist=eqlist, restrictions=constraints,
                  method="2wayWB", data=datap)
R> printSUR(mod5)

SUR estimation results
Method: Two-way WB

Unbalanced Panel: n = 14288, T = 1-10, N = 34140

============= ============== ============ =========== ========
 Coefficient     Estimate      Std.Error    t-value   p-value
============= ============== ============ =========== ========
...
P4NT2W1           -0.53298      0.15974    -3.33652   0.00084
...
P4NT2W2           -0.54869      0.16763    -3.27317   0.00106
...
P4NT2W3           -0.75340      0.27122    -2.77785   0.00548
...
const             -0.83051      0.01417   -58.59085   0.00000
P4NT2W4            6.61559      0.37372    17.70190   0.00000
P5NT2W4           -0.30735      0.16156    -1.90236   0.05714
P6NT2W4          -15.08214      1.68265    -8.96331   0.00000
A1NT2W4           -0.34579      0.05896    -5.86480   0.00000
A2NT2W4            0.00351      0.04708     0.07451   0.94060
A3NT2W4            0.17160      0.07671     2.23700   0.02530
A4NT2W4            0.22384      0.03552     6.30180   0.00000
VARPPNT2W4_1      -0.00050      0.00075    -0.67566   0.49926
VARPPNT2W4_2       0.00008      0.00001     6.33139   0.00000
VARPPNT2W4_3       0.00393      0.00081     4.82163   0.00000
COVPPNT2W4_1_2    -0.00876      0.00223    -3.92041   0.00008
COVPPNT2W4_1_3     0.00282      0.00035     8.06623   0.00000
COVPPNT2W4_2_3    -0.00324      0.00042    -7.67224   0.00000
ST2W4              0.10052      0.00229    43.90089   0.00000
ET2W4             -6.49535      2.02316    -3.21049   0.00132
ZT2W4              0.00051      0.00010     5.07995   0.00000
NOR2W4           117.08097      1.91326    61.19441   0.00000
...
============= ============== ============ =========== ========
```

```
R> mod6 <- SURest(eqlist=eqlist, restrictions=constraints,
                  method="2wayQUE", data=datap)
R> printSUR(mod6)

SUR estimation results
Method: Two-way QUE

Unbalanced Panel: n = 14288, T = 1-10, N = 34140

=============  ==============  ============  ===========  ========
 Coefficient      Estimate       Std.Error     t-value    p-value
=============  ==============  ============  ===========  ========
...
P4NT2W1              -0.53046       0.15839     -3.34915   0.00082
...
P4NT2W2              -0.53467       0.16465     -3.24728   0.00116
...
P4NT2W3              -0.80951       0.26664     -3.03603   0.00240
...
const                -0.83719       0.01408    -59.46964   0.00000
P4NT2W4               6.60235       0.36657     18.01121   0.00000
P5NT2W4              -0.27386       0.15816     -1.73154   0.08336
P6NT2W4             -15.06467       1.64455     -9.16034   0.00000
A1NT2W4              -0.33408       0.05768     -5.79170   0.00000
A2NT2W4               0.00615       0.04568      0.13472   0.89284
A3NT2W4               0.12347       0.07530      1.63970   0.10108
A4NT2W4               0.22817       0.03470      6.57575   0.00000
VARPPNT2W4_1         -0.00056       0.00073     -0.77620   0.43764
VARPPNT2W4_2          0.00009       0.00001      6.50291   0.00000
VARPPNT2W4_3          0.00379       0.00080      4.75555   0.00000
COVPPNT2W4_1_2       -0.00868       0.00218     -3.97638   0.00008
COVPPNT2W4_1_3        0.00281       0.00034      8.22231   0.00000
COVPPNT2W4_2_3       -0.00325       0.00041     -7.89088   0.00000
ST2W4                 0.10033       0.00227     44.23391   0.00000
ET2W4                -5.78553       1.98390     -2.91624   0.00354
ZT2W4                 0.00052       0.00010      5.25244   0.00000
NOR2W4              118.01657       1.89881     62.15288   0.00000
...
=============  ==============  ============  ===========  ========
```

4. SUMMARY AND DISCUSSION

In this paper we introduced the **panelSUR** package for the R environment. The **panelSUR** package provides the possibility to estimate systems of linear equations by different methods, such as the one-way EC model estimation method suggested by Biørn (2004) and the two-way EC model estimation methods suggested by Platoni et al. (2012a) (i.e., the QUE and WB procedures). These are generalized least squares (GLS) estimators that take into account the covariance structure of the residuals, including individual-specific and time-specific effects.

Although other statistical and econometric software packages can usually manage systems of linear equations, the **panelSUR** package distinguishes itself by efficiently addressing Seemingly Unrelated Regression (SUR) systems in unbalanced panel data. Indeed, some other software do not effectively handle SUR systems in panel data; instead, they either estimate a single equation for all individuals in the panel or use pooled data to estimate multiple equations. Moreover, the Stata module **xtsur**, although capable of estimating SUR system in panel data, is based on the methodology originally developed by Biørn (2004) and thus it merely estimates the one-way case.

The **panelSUR** package is able to take into account equation restrictions involving equality between two parameters as well. The package also includes an illustrative simulated unbalanced panel dataset.

For the methods developed in the current version of the package, the disturbances of the individual equations are assumed to be independent and identically distributed (i.i.d.). Future work could be done enhancing the package by the inclusion of methods taking into account also heteroscedastic disturbances, on the trace paved by Platoni et al. (2012a), and extending the capability to accomodate restrictions in the form of linear combinations involving more than two parameters.

A. EXPECTED VALUES AND COVARIANCE MATRICES OF $\mathbf{U_{IT}}$, $\boldsymbol{\mu_I}$, AND $\boldsymbol{\nu_T}$

If we do not have cross-equation restrictions, we can assume $\mathrm{E}(u_{mit}|\mathrm{x}_{1it}, \mathrm{x}_{2it}, \ldots, \mathrm{x}_{Mit}) = 0$, and then $\mathrm{E}(y_{mit}|\mathrm{x}_{1it}, \mathrm{x}_{2it}, \ldots, \mathrm{x}_{Mit}) = \mathrm{E}(y_{mit}|\mathrm{x}_{mit}) = \mathrm{x}_{mit}\boldsymbol{\beta}_m$. On the contrary, if we have cross-equation restrictions we can only assume $\mathrm{E}(\mathrm{u}_{it}|\mathrm{x}_{it}) = 0$, where $\mathrm{u}_{it} \equiv (u_{1it}, \ldots, u_{Mit})^\top$ and $\mathrm{x}_{it} \equiv (\mathrm{x}_{1it}, \mathrm{x}_{2it}, \ldots, \mathrm{x}_{Mit})$. With the $M \times 1$ vectors $\boldsymbol{\mu}_i \equiv (\mu_{1i}, \ldots, \mu_{Mi})^\top$ and $\boldsymbol{\nu}_t \equiv (\nu_{1t}, \ldots, \nu_{Mt})^\top$, we assume

$$
\begin{aligned}
&\mathrm{E}\left(\mu_{mi}, \mu_{jh}\right) && \begin{cases} = \sigma^2_{\mu_{mj}} & i = h \\ = 0 & i \neq h, \end{cases} \\
&\mathrm{E}\left(\nu_{mt}, \nu_{js}\right) && \begin{cases} = \sigma^2_{\nu_{mj}} & t = s \\ = 0 & t \neq s, \end{cases} \\
&\mathrm{E}\left(u_{mit}, u_{jhs}\right) && \begin{cases} = \sigma^2_{u_{mj}} & i = h \text{ and } t = s \\ = 0 & i \neq h \text{ and/or } t \neq s, \end{cases}
\end{aligned}
\tag{25}
$$

and then $\tilde{\boldsymbol{\mu}}_m \equiv (\mu_{m1}, \ldots, \mu_{mN})^\top$, $\tilde{\boldsymbol{\nu}}_m \equiv (\nu_{m1}, \ldots, \nu_{mT})^\top$, and $\mathrm{u}_m \equiv (u_{m11}, u_{m12}, \ldots, u_{m1T_1}, u_{m21}, \ldots, u_{mNT_N})^\top$ are respectively $N \times 1$, $T \times 1$, and $n \times 1$ random vectors with zero means and covariance matrix

$$
\mathrm{E}\left(\begin{pmatrix} \tilde{\boldsymbol{\mu}}_m \\ \tilde{\boldsymbol{\nu}}_m \\ \mathrm{u}_m \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\mu}}_j^\top & \tilde{\boldsymbol{\nu}}_j^\top & \mathrm{u}_j^\top \end{pmatrix}\right) = \begin{bmatrix} \sigma^2_{\mu_{mj}} & 0 & 0 \\ 0 & \sigma^2_{\nu_{mj}} & 0 \\ 0 & 0 & \sigma^2_{u_{mj}} \end{bmatrix}. \tag{26}
$$

With the $NM \times 1$ vector $\boldsymbol{\mu} \equiv (\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_N^\top)^\top$, the $TM \times 1$ vector $\boldsymbol{\nu} \equiv (\boldsymbol{\nu}_1^\top, \ldots, \boldsymbol{\nu}_T^\top)^\top$, and the $nM \times 1$ vector $\mathrm{u} \equiv (\mathrm{u}_{11}^\top, \mathrm{u}_{12}^\top, \ldots, \mathrm{u}_{1T_1}^\top, \mathrm{u}_{21}^\top, \ldots, \mathrm{u}_{NT_N}^\top)^\top$, since we have $\boldsymbol{\mu} \sim (0, \Sigma_\mu)$, $\boldsymbol{\nu} \sim (0, \Sigma_\nu)$, and $\mathrm{u} \sim (0, \Sigma_u)$, with the $M \times M$ matrices $\Sigma_\mu = [\sigma^2_{\mu_{mj}}]$, $\Sigma_\nu = [\sigma^2_{\nu_{mj}}]$, and $\Sigma_u = [\sigma^2_{u_{mj}}]$, we can assume that the expected values of the $M \times 1$ vectors μ_i, ν_t, and u_{it} are zero and their covariance matrices are equal to Σ_μ, Σ_ν, and Σ_u. It follows that $E(\varepsilon_{it}\varepsilon_{hs}^\top) = \delta_{ih}\Sigma_\mu + \delta_{ts}\Sigma_\nu + \delta_{ih}\delta_{ts}\Sigma_u$, with $\delta_{ih} = 1$ for $i = h$ and $\delta_{ih} = 0$ for $i \neq h$, $\delta_{ts} = 1$ for $t = s$, and $\delta_{ts} = 0$ for $t \neq s$.

REFERENCES

Baltagi, B. H. (1980). On seemingly unrelated regressions with error components. *Econometrica* 48(6):1547-1551.

Baltagi, B. H. (1981). Pooling: An experimental study of alternative testing and estimation procedures in a two-way error component model. *Journal of Econometrics* 17(1):21-49.

Biørn, E. (2004). Regression systems for unbalanced panel data: A stepwise maximum likelihood procedure. *Journal of Econometrics* 122(2):281-291.

Croissant Y., Millo G. (2008). Panel Data Econometrics in R: The plm Package. *of Statistical Software* 27(2):1-43.

Davis, P. (2002). Estimating multi-way error components models with unbalanced data structures. *Journal of Econometrics* 106(1):67-95.

Magnus, J. R. (1982). Multivariate error components analysis of linear and non-linear regression models by maximum likelihood. *Journal of Econometrics* 19(2-3):239-285.

Nerlove, M. (1971). Further evidence on the estimation of dynamic economic relations from a time series of cross sections. *Econometrica* 39(2):359-382.

Platoni, S., Barbieri, L., Sckokai, P., Moro, D. (2020). Heteroscedastic Stratified Two-way EC Models of Single Equations and SUR Systems. *Econometrics and Statistics* 15:46-66.

Platoni, S., Sckokai, P., Moro, D. (2012). A note on two-way ECM estimation of SUR systems on unbalanced panel data. *Econometric Reviews* 31(2):119-141.

Platoni, S., Sckokai, P., Moro, D. (2012). Panel Data Estimation Techniques and Farm-level Data Models. *American Journal of Agricultural Economics* 94(5):1202-1271.

R Core Team (2022). R: A Language and Environment for Statistical Computing. URL `https://www.R-project.org/`.

Shonkwiler. J.S., Steven, T.Y. (1999). Two-Step Estimation of a Censored System of Equations. *American Journal of Agricultural Economics* 81(4):972-982.

Tauchmann, H. (2005). Efficiency of two-step estimators for censored systems of equations: Shonkwiler and Yen reconsidered. *Applied Economics* 37(4):367-374.

Wansbeek, T., Kapteyn, A. (1989). Estimation of the error-components model with incomplete panels. *Journal of Econometrics* 41(3):341-361.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias, *Journal of the American statistical Association* 57(298):348-368.

Printed by
Gi&Gi srl - Triuggio (MB)
June 2024

9788834358429