Revista Colombiana de Estadística
January 2018, Volume 41, Issue 1, pp. 31 to 52
DOI: http://dx.doi.org/10.15446/rce.v41n1.61885

Estimating dynamic Panel data. A practical approach to perform long panels

Estimando Datos de panel dinámicos. Un enfoque práctico para abordar paneles largos

Romilio Labra[1,a], Celia Torrecillas[2,b]

[1] Instituto de Innovación Basada en Ciencia, Universidad de Talca, Talca, Chile
[2] Departamento de Administración y Dirección de Empresas, Facultad de Ciencias Sociales y de la Comunicación, Universidad Europea de Madrid, Madrid, España

Abstract

Panel data methodology is one of the most popular tools for quantitative analysis in the field of social sciences, particularly on topics related to economics and business. This technique allows simultaneously addressing individual effects, numerous periods, and in turn, the endogeneity of the model or independent regressors. Despite these advantages, there are several methodological and practical limitations to perform estimations using this tool. There are two types of models that can be estimated with Panel data: Static and Dynamic, the former is the most developed while dynamic models still have some theoretical and practical constraints. This paper focuses precisely on the latter, Dynamic panel data, using an approach that combines theory and praxis, and paying special attention on its applicability on macroeonomic data, specially datasets with a long period of time and a small number of individuals, also called long panels.

***Key words*:** Dynamic Panels; Endogenous Models; Overidentification; Panel Data; Stata; xtabond2.

Resumen

La metodología de Datos de Panel es una de las técnicas más usadas para realizar análisis cuantitativos en el ámbito de las ciencias sociales, especialmente en temas relacionados con la economía y los negocios. Su riqueza reside en que esta técnica permite trabajar con varios periodos de tiempo, incorporar los efectos individuales, y a su vez, tratar la endogeneidad. A pesar de estas ventajas, existen diversos obstáculos para su implementación,

[a] PhD. E-mail: ernesto.labra@utalca.cl
[b] PhD. E-mail: celiatorrecillas@gmail.com

tanto metodológicos como operativos. Dentro de los tipos de modelos que se pueden estimar con Datos de Panel, los de carácter estáticos han sido los más desarrollados, persistiendo aún carencias teórico-prácticas para los modelos dinámicos. Este artículo pone precisamente su énfasis en estos últimos, aplicando un enfoque que conjuga la teoría y la praxis, y prestando especial atención a su aplicabilidad para datos macroeconómicos, fundamentalmente para paneles que poseen un período de tiempo largo y un número de individuos pequeño.

***Palabras clave*:** datos de panel; datos de panel dinámicos; modelos endógenos; sobreidentificación; stata; xtabond2.

1. Introduction

Studies on Panel data methodology began in the XIX answering new questions that Pool data analysis or Time series could not directly solve. The first works on this methodology was focused on lineal regressions and static models, where fix and random effects were determined assuming a fixed temporal effect without paying enough attention to endogenous relationships. To analyse these interactions, a new tool was developed in the XX: Dynamic models; Balestra & Nerlove (1966), Nerlove (1971), Maddala (1971, 1975) are some of the first works. Finally, in the 70's, empirical studies on dynamic panel data began to be published in specialized journals.

Dynamic Panel data methodology offers some advantages in comparison to the Static version. The possibility to address the heterogeneity of the individuals and also the use of several instrumental variables in order to deal with the endogeneity of the variables of the model, also known as "lagged variables". Moreover, along with the estimation of models with endogenous variables, it is possible to perform more sophisticated models (Ruíz-Porras 2012). However, dynamic panel data also has some weaknesses. First, estimators can be unstable and the reported values could depend on characteristics of the sample. Also, the use of lagged variables not necessarily can deal with serial correlation problems (Pérez-López 2008). In addition, it is complex to find appropriately instruments to some endogenous regressors when only weak instruments are available. Nevertheless, one of the main limitations of this methodology is the analysis for long time periods (long t) and few individuals (short n), which could result in the overidentification of the model (Ruíz-Porras 2012).

Several empirical studies on the field of Economy are using databases with long time periods and small number of individuals, for example when researcher try to understand the effect of key factors on performance of companies, industries or territories. In order to built these models, previous authors have proposed to treat the equations from the different cross section units as a system of seemingly unrelated regression equations (SURE) and then estimate the system by generalized least squares (GLS) techniques (Pesaran 2006), while others assume Panel data as a more adequate methodology to deal with these estimations. This fact is the main target of this article, which provides some alternatives to face this situation and estimate dynamic models with long panels (long t and short n).

During the nineties, studies of endogenous models using dynamic panel data (DPD) were usual and some works on this methodology were carried out. Relevant contributions on DPD by Arellano & Bond (1991), Arellano & Bover (1995), Blundell & Bond (1998) and Roodman (2009) were provided in order to improve the understanding of the complex economic processes by empirical researches. Although, it has been more than thirty years from the first works, this technique still has some open questions. Thus, the purpose of this article is to guide the reader in the use of dynamic panel data and provide some clues to solve limitations when panels are formed by a long t and short n. This restriction is addressed by using Stata, and solutions are offered in this text.

The paper is made up of four sections. The next section offers a review on Dynamic Panel Data including the models to be analyzed by using this methodology. Then, a detailed description about how estimate long panels is included. The fourth section includes examples of endogenous model estimates using Stata. Finally, remarked conclusions are provided.

2. Review on Dynamic Panel Data

2.1. Evolution and Advance on Panel Data Methodology

In the last fifty years, Panel data methodology has become one of the most popular tools for empirical studies in different fields of knowledge. There has been an important progress in the knowledge for static models, but in the dynamic version still remain some theoretical and practical constraints. The purpose of this paper is to provide some clues and recommendations for the use of dynamic panel data, specifically for the performing of endogenous models with long panels. Panel data is a statistical tool to perform models using a number of individuals (companies, countries, households, etc.) across a defined period of time. This technique differs from cross-sectional analysis, which is used to perform an analysis of several individuals at a specific point in time, and the methodology of time series, which corresponds to the analysis of the same individual across time. Thus, the use of panel data requires two conditions: data from different individuals (n) collected over time (t). In addition to these conditions, restrictions may also arise due to the number of observations and the relationship between n and t. The recommendation to perform a model with panel data is to use a large number of individuals (n) and a small period of time (t), in order to have adequate degrees of freedom and avoid overidentification.

This methodology has been used more frequently in studies at firm level, because databases usually have a large number (n) of observations in a short period of time (t). This condition offers the advantage of capturing the variability of the phenomenon, through observation of a large number of cases. At an aggregate level (e.g.: countries, regions, sectors, etc.), whose databases frequently have a small n/t relationship, even less than 1, some serious difficulties arise when studies of endogenous models are carried out. Figure 1 shows an example comparing OLS and Panel analysis where the individuals effect has been taken into account providing a better adjustment and thus, improving the explanatory capacity. Different results (models) are obtained in the example when the models are performed by OLS or Panel. The above is a consequence of individual effects, which can be assumed by panel data methodology. In fact, individual effects (dashed line) generate a greater slope of the function than OLS estimating (solid line) and better adjusting the model to the observed data, improving the explanatory capacity.

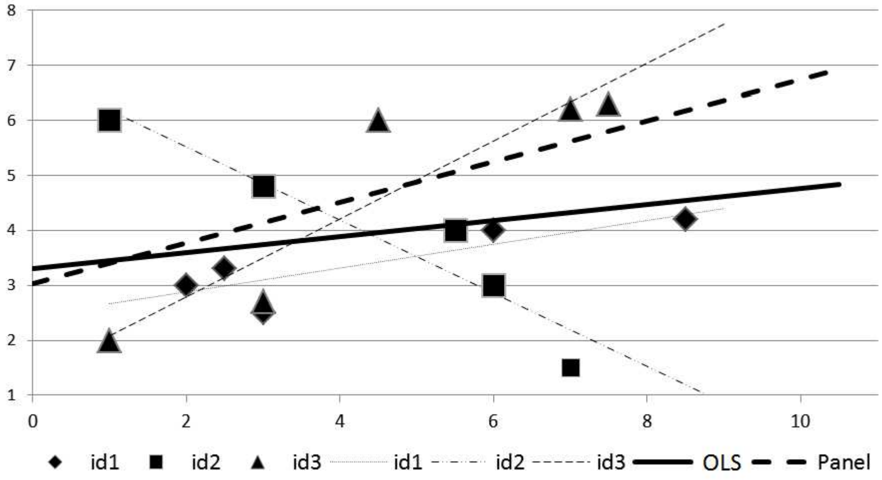

Figure 1: Model estimates by OLS and Panel data methodologies.

As it has been mentioned above, there are two main types of Panel data. Static panels, used to estimate static models, and Dynamic panels, more suitable to perform endogenous models. Static panels can be classified into models with fixed or random effects, depending on how they consider the individual effects, assuming in both cases these effects as constant over time. The above restriction makes static models limited to consider the dynamics of time-varying, or the endogeneity. On the contrary, dynamic panel data allow us to treat endogeneity of variables and model.

From an evolutionary perspective, Nelson & Winter (1982) and Dosi (1988) indicated that endogenous models are highly dependent on the past and its accumulative process. Dynamic panels allow including an endogenous structure into the model through instrumental variables. This endogeneity is defined as the existence of correlation between the dependent variable and the error term, which is related to the causal relationship between the variables explainin the model (Mileva 2007, Wooldridge 2013), inadequate data quality, autoregression and autocorrelated errors and/or omission of relevant variables. In economic terms, endogeneity can be interpreted as the effect of the past on the present, both on the model (dependent variable) and on the independent variables, or as the causality relationship between regressors and explained variable along the time.

The inclusion of the dependent variable as regressor, consistent with the work reported by classical authors such as Arellano & Bond (1991), Arellano & Bover (1995) and Blundell & Bond (1998), is performed by using lagged endogenous terms as a way to avoid the correlation problems between variables, defining (Y) : $Y_{(it-n)}$.

The second term of the function (regressors) corresponds to the lag of dependent variable $(Y_{(it-n)})$ plus the independent variables (X_{it}). Due to the causality is related to time, the regressor is included as the lag of $Y_{(it)}$

$$Y_{it} = \alpha Y_{it-n} + \beta_i X_{it} + \omega_{it} \tag{1}$$

Where:
Y_{it}: dependent variable of individuals i in time t
Y_{it-n}: lag of dependent variable. Individuals i en time $t-1$
α : constant
β_i : coefficient of variable i
X_{it}: independent variable i in time t
$\omega_{it} : \varepsilon_i + \mu_{it}$

In addition, not only the lagged variables can be used as instruments of endogenous variables, but also others independent variables correlated to the regressor target but not correlated to the error term of the model. In general, these types of instruments are not very easy to detect, and many times they can be not completely correlated to the endogenous variable.

2.2. Types of Dynamic Panel data

The evolution in the analysis of dynamic Panel data and the building of estimators, have introduced new possibilities for the analysis of endogenous models. These models have been specially focused on the econometric analysis of the endogeneity.

Two main ways have been developed to address the endogeneity in the models, in addition to traditional instrumental variables; the first one, is to build instrumental variables in levels, while a second choice corresponds to the generation of those variables but in differences. However, even when the literature has showed advances in these analysis, there are some difficulties in the application of the dynamic panel data. This is particularly discussed in this document.

2.2.1. Dynamic Panel Functions: Instrumental Variables in Differences and Levels

The first method for the treatment of the endogeneity problem uses the instrumental variables obtained through lags of the endogenous variables. Depending on the estimator used, these lags may be applied in differences or levels. The differences between both methods will be expressed in the following equations:

Instrument in differences:

$$X_{(t-n)} - X_{(t-(n-1))} \tag{2}$$

Instrument in levels:

$$X_{(t-n)} \tag{3}$$

The use of instruments in differences or levels is expressed in the following equations:

Equations in differences:

$$\Delta Y_{t-1} = Y_{t-2} - Y_{t-1} \tag{4}$$

Equations in levels:

$$Y_t = Y_{t-1}; Y_{t-(n-1)} = Y_{t-n} \tag{5}$$

Where, Y_{t-n} is the instrument of $Y_{t-(n-1)}$

Considering the building of instruments in dynamic panel data, it is possible to find different estimators:

The first one was developed by Arellano and Bond in 1991 (Arellano & Bond 1991). It is known as Difference GMM, because this estimator uses as instruments the lags in differences.

Latter, it was developed the estimator that uses as instrumental variables the lags in differences and levels. This change allowed to work with panel data composed by a small period of time, and therefore with a small number of instruments. It is known as System GMM and it was developed by (Arellano & Bover 1995).

A third estimator was developed by Roodman (2006). It is called xtabond2. This estimator follows the same logic that System GMM, but it introduces more options in the used of the instruments. In addition, xtabond2 allows us to work separately the endogeneity of the dependent or independent variables.

As we have mentioned, System GMM uses the instruments in level and differences. The equations that allow its calculation are as follows:

Equations in differences and levels. System GMM

$$Y_{it} = \alpha_{i,t-1} + \beta X_{it} + \varepsilon_{it} \tag{6}$$

$$\varepsilon_{it} = \mu_i + \vartheta_{it} \tag{7}$$

$$E(u_i) = E(\vartheta_{it}) = E(\mu_i \vartheta_{it}) = 0 \tag{8}$$

Y_{it} is the dependent variable of i (individual) in t (period of time)

X_{it} is the independent variable of i (individual) in t (period of time).

The error term ε_{it} has two orthogonal components:

$$\mu_i = \text{fixed effects}$$
$$\vartheta_{it} = \text{idiosyncratic shocks}$$

The use of dynamic panel data in differences (Difference GMM) and system (System GMM) requires different commands in Stata:

xtbond (Arellano & Bond 1991). This command uses as instrumental variable the lags of endogenous variable in differences (Difference GMM).

xtdpdsys (Arellano & Bover 1995). This command uses as instrumental variables of endogenous variable the lags in differences and levels (Difference and System GMM)

xtabond2 (Roodman 2006). Similarly to xtdpdsys, it uses the instrumental variables of endogenous variable as lags in levels and differences. This is not an official command in Stata, but it is an option given by Roodman (2006).

xtdpd. It is used for the regression of endogenous variables as instruments in differences or levels. According to Cameron & Trivedi (2009), the use of this command will allow to correct the model of the average moving, being detected by the Arellano and Bond Test (Autocorrelation of second order).

In addition, the estimators mentioned above, allow us to do the analysis through two alternatives: One step and Two steps, depending on if the weight matrix is homocedastic or heterocedastic. Literature indicates that Two steps estimators are more efficient; therefore it is recommendable the use of the heterocedastic matrix in this type of estimations.

One step: It uses only the homocedastic weight matrix for the estimation. Two steps: It uses the heterocedastic weight matrix for the estimation.

The differentiation between these alternatives is the key for the determination of overidentification in a dynamic model, as we will analyzed in the next section.

2.2.2. Main Issues in the Estimation of Dynamic Panel data Using GMM

The utilization of GMM in the estimation has two main issues: the proliferation of instruments and the serial autocorrelation of errors. These two issues will be higher when the panel used is made up by a sample with a big period of time and reduced number of individuals.

The proliferation of instruments refers to the existence of a higher level of instruments. This will cause overidentification in the model as a consequence of the generation of instrumental variables in differences and levels (with the exception of the Arellano and Bond estimator that only uses as instrumental variables lags in differences). In order to check if the number of instruments is adequated and it doesn't produce overidenfitication, there are two tests available: Sargan test and Hansen test (both tests are explained in section 4.4) (Sargan 1958).

The dynamic panel data requires that the error cannot be serially correlated: This condition of the serial autocorrelation of errors can be avoided using the Arellano and Bond test (this is explained in section 4.4.3).

3. Estimating Dynamic Models Using Macro Data and Long Panels

Panel data are part of the techniques available to perform models from databases formed by a small number of individuals observed in a long time. However, the length of T (time) and N (individuals) could change according to the research analysis which could produce uncorrected results in the estimations. Therefore in order to deal with the length of N and T, some authors have pointed out some ideas that can help in dealing with these differences. In what follows we present approaches for the following scenarios: 1) N small, T large; 2) N and T large; 3) N small, T large.

Previous authors propose to treat the equations from the different cross section units as a system of seemingly unrelated regression equations (SURE) and then estimate the system by generalized least squares (GLS) techniques (Pesaran 2006). Specifically when N is small relative to T and the error are uncorrelated with the regressors crossection dependence can be modelled using SURE (Chudik, Pesaran & Tosetti 2011).

Although Pesaran (2006) pointed out some ideas for dealing with different size of N and T, the proposal is not totally appropriate when both N and T are large, as it is the case of countries studies. For N and T large some authors propose restricting the covariance matrix of the error using a common factor specification with a fixed number of unobserved factors (Hoechile 1933, Phillips & Sul 2003). However, some econometric error occurs when N is large. For that reason, Pesaran (2006) proposed to apply the Common Correlated Effect estimators (CCE) when N and T tend to infinite.

However, is quite common to have a small N and a large T. In this case, four solutions have been proposed; 1)Running a separate regression for each group and averaging the coefficient over groups. 2) Combine the data defining a common slope, allowing for fixed or random intercepts and estimating pooled regressions (Mairesse & Griliches 1988). 3) Take the data average over group and estimate the aggregate time series regressions (Pesaran, Pierse & Kumar 1989, Lee, Pesaran & Pierse 1990). Finally, 4) Averaging the data over time and estimating cross section regression on group means (Barro 1991). The solutions mentioned above present some limitations. The group mean estimator obtained by the average of the coefficients for each group is consistent for large N and T, but the pooled and aggregate estimators are not consistent in dynamic models and there is a bias (Pesaran & Smith 1995).

Another approach deals with databases formed by small N and large T using Panel data methodology as solution to this condition (N small in comparison to T). This work is along this line of research.

The use of large number of individuals and short period of time is the most common type of data in dynamic panel analysis. It is called "*Short panels*". The literature does not specify a number of individuals (n) or time (t) to classify the panels as long or short. However some authors have indicated the following rule: a suitable n could be greater than 100, while the t should not exceed 15 periods, and ideally it should be less than 10, if the target is to estimate dynamic models with panel data (Roodman 2009). This is the case of several studies based on databases compiled from surveys such as CIS (Community Innovation Survey), PITEC (Panel of Technological Innovation, Spain), national surveys of innovation and others databases of companies and citizens provided by national and international organizations.

When we try to estimate dynamic models with panels conformed by n relatively small ($n < 100$) and t large ($t > 15$), using lags of variables as instruments of endogenous terms, we find additional difficulties due to the panel data structure and the way of instrumental variable generation. This fact is caused by the incorporation of lags of endogenous variable as its instrument(s), which must be correlated to the endogenous regressor and $E(\mu|x) = 0$. This alternative of instruments (lags of endogenous variables) resolves the problem finding a suitable instrument to endogenous regressors.

Anderson & Hsiao (1981), Arellano & Bond (1991) and Arellano & Bover (1995) have demonstrated the importance of lags as instruments, and the relevance to estimate dynamic models. Nevertheless, when using long panels an important obstacle emerges: the proliferation of instruments (Roodman 2009). This is because the number of instruments to be generated is directly related to the length of the panel (number of periods). For example, for a variable with $t = 5$, the number of potential instruments is 12 (from equations in differences and 3 from equation in levels) when we use GMM methodology.

In the case of equations in differences, the number of instruments is defined as follows:

Function:

$$\Delta Y_{it} = \delta \Delta Y_{i(t-i)} + \Delta\varepsilon_{it} \tag{9}$$

Instruments:

$$Y_{i1}, Y_{i2}, \ldots\ldots, Y_{it-2} \tag{10}$$

If there are one or more endogenous variables, the number of instruments increases even more, as each regressor is instrumentalized by all their differences and levels (with GMM). This proliferation of instruments, initially was seen as favorable, since it increased the efficiency of the estimator (Arellano & Bond 1991), however, it causes an "*overidentification*" of the model, mainly when the number of degrees of freedom is small, e.g. when there are few individuals. Therefore, as the panel grows in periods and decreases on number of individuals, the probability of overidentification increases. In Table 1 we show a brief summary of the main problems that arise with the use of panel analysis.

Table 1: Main obstacles found in the estimation of DPD considering the number of individuals (n) and period (t).

		Number of individuals (n)	
		High	Low
Period of time (t)	High	Low probability of overidentification	High probability of overidentification
	Low	Normal condition of panel	The number of observations can be insufficient to perform the model

In order to solve the overidentification problem, Roodman (2009) conducted a detailed analysis and proposed mechanisms to adequately test the existence of excess of instruments, through the Sargan and Hansen tests. According to the author, Sargan test is adequate when the estimation is performed considering an homoscedastic weight matrix, as is the case of the *One step option*. With Stata, the command to run this test is *estat Sargan* and it is available as a model postestimation.

The null hypothesis:

$H_0 =$ overidentification restrictions apply

Meanwhile, the Hansen test detects overidentification in presence of an heterocedastic matrix. This is the case when using the *Two step* and *vce(robust)* options. This test is directly reported when it is used the estimator *xtabond2* in Stata.

The null hypothesis of this test the same as the Sargan test, because both are identifying the existence of instrument excesses.

$H_0 =$ overidentification restrictions apply.

In order to avoid an overidentification of the model, the number of individuals or groups must be greater than the number of instruments used. Therefore, reducing the number of instruments becomes a necessary condition when we use long panels. The literature shows various alternatives to solve this problem depending on the nature of the model, the purpose of the analysis, the length of the panel and the characteristics of the variables. The first alternative is to reduce t, dividing the analysis into two sections (two separate models). Other possibility is to group the periods (e.g. using biennia, trienniums or others). However, these options are limited, because they reduce the information available for the analysis, affecting the variance.

Another alternative is to reduce the instruments through the restriction of lags. As an endogenous model is specified incorporating the lag(s) of the independent variable (Y) as regressor(s), it is common to limit to one or two the periods (lags), that is $Y_{(t-1)}$ and $Y_{(t-2)}$ (commonly known as L1 and L2, respectively). If we suspect of a delayed (endogenous) effect, it is recommended to add more lags, a situation that can be offset by the elimination of those closest to t_0 due to each L (lag of each endogenous regressors) incorporates more instruments.

It is also possible to reduce the generation of instrumental variables, either lag of Y or endogenous regressors, using only equations in differences or levels. In addition, if we need to further reduce these types of variables, we can restrict lags of each variable to a value between $t-1$ and t_n, in other words, to estimate the model using as instruments only those generated for a time interval and not for the entire period of the panel.

In order to select the option to reduce the number of instrument, some criteria are proposed:

- Sample characteristics.
    - ◇ Number of individuals (n)
    - ◇ Time periods (t)
- Literature review on characteristics of model (endogenous or not) and the regressors.
- Serial correlation between model's errors
- Overidentification. In order to manage many instruments it is required more than one alternative to limit them.

In endogenous models, in addition to the overidentification discussed above, additional drawbacks related to the serial second-order autocorrelation of residues can arise, indicating that the instrument used is not consistent. Given this limitation, we constantly need to test instrument variables in order to define the most appropriate regressor, because even when the number is suitable, can remain the serial autocorrelation inconvenient. To identify whether or not autocorrelation, Arellano and Bond test should be used, as follows:

Arellano and Bond test.

The null hypothesis is:

- Ho: There not exist autocorrelation.
- Stata by default delivers results for the order 1 and 2 (Ar (1) and Ar (2)). When Arellano and Bond test indicates that there is serial correlation in both levels, probably we are facing a unit root model.

Finally, we show some examples of applications where Panel data was used with small N and large T: in: Roodman (2009), Labra & Torrecillas (2014), Álvarez & Labra (2014), Torrecillas, Fischer & Sánchez (2017), Santos-Arteaga, Torrecillas & Tavana (2017).

4. Modeling Endogenous Functions With Panel Data: Step by Step[1]

This section contains the syntax for the use of dynamic panel data in Stata and the interpretation of the set of estimators: *xtabond*, *xtdpdsys* and *xtabond2*.

4.1. xtabond Estimators (Instrumental Variables used in Differences)

To perform a regression using *xtabond* we will distinguish between models with endogenous, exogenous and/or predetermined variables.

4.1.1. Models with exogenous independent variables.

Step 1

xtabond vardep var1 var2 var3 varn, lags(#) twostep.
estat sargan

Step 2

xtabond vardep var1 var2 var3, varn, lags(#) vce(robust) twostep.
estat abond.

Where, var1, var2, var3 and var n, are independent exogenous variables.

Firstly, we should do the estimation without the *vce(robust)* option, and then apply the Sargan test (this test only works without this option).

The order in the syntax will be as follows: xtabond indicates to Stata that you are using dynamic panel data, then, the dependent variable (vardep) has to be written, and later the independent exogenous variables (in the example: var1, var2, var3 and var n). Finally, after the comma and with the expression lags, you should introduce the number of lags of the dependent variable as regressor.

Secondly, the model is estimated with the option vce(robust). After this option we will add the Arellano and Bond test to determine the existence of serial autocorrelation or not (estat abond).

4.1.2. Models with Predetermined Independent Variables

Step 1

xtabond vardepend var1 var2 var3, lags(#)twostep pre (var4, va5, lagstructur(#,#)) estat sargan

Step 2

xtabond vardepend var1 var2 var3, lags(#) twostep vce(robust) pre(var4, var5, lagstructur(#,#)) estat abond

[1] Some examples for the application of this model with Stata are included in Labra and Torrecillas (2014)

Where, var1, var2 y var3, are exogenous independent variables and var4, var5 are predetermined independent variables, which is indicated with the following expression: pre(var4 y var5).

In order to indicate to Stata the use of independent variables as predetermined, we use the following syntax after the comma: pre (*var4, var5, lagstructur(#,#)*). Inside the parentheses we will introduce the predetermined variables (in the examples var4 y var5) and the limitations of the lags (*lagstructur(#,#)*). The first # indicates the number of lags introduced in the model, and the second # indicates the maximum quantity of lags.

4.1.3. Models With Endogenous Independent Variables

Step 1

xtabond vardepend var1 var2 var3, lags(#)twostep endog (var6, va7, lagstructur(#,#))
estat sargan

Step 2

xtabond vardepend var1 var2 var3, lags(#) twostep vce(robust)
endog(var6, var7, lagstructur(#,#)) estat abond

Where: *var1, var2* y *var3* are exogenous independent variables and var6 and var7 are independent endogenous variables (endog(var6, var7)) In order to indicate to Stata that the variables are endogenous we will use the following syntax after the comma: endog(*var6, var7, lagstructur(#,#)*). Inside the parentheses we will introduce the endogenous variables and the limitations for the lags.

4.1.4. Other Combinations of the Variables

It is possible to combine different types of independent variables: exogenous, predetermined and/or endogenous.

xtabond vardepend var1 var2 var3, lags(#) twostep vce(robust) pre(var4, var5, lagstructur(#,#)) endog(var6, var7, lagstructur(#,#))

The limitation of the (lags) can be specified for each variable or groups of variables.

xtabond vardepend var1 var2 var3, lags(#) twostep vce(robust) endog(var6, var7, lagstructur(1,.)) endog(var8, lagstructur(2,2))

Where, *var1, var2* and *var3* are the independent exogenous variables and var6, var7 and var8 are the independent endogenous variables.

The number maximum of lags will depend on the period of time of the sample, taking into account that when it is used, the estimator in difference, one t is lost for each difference.

In the estimation other commands can be used to specify the limitation of the instruments:

1. Maxdelp (#) Maximum number of lags of the dependent variables that can be used as instruments.

2. Maxlags (#) Maximum number of lags of the predetermined and endogenous variables.

4.2. xtdpdsys Estimators (Instruments in Differences and Levels)

The syntax for the using of this command is:

xtdpdsys vardep var1 var2, lags(#) twostep vce(robust) pre(var4, var5, lagstructur(#,#)) endog(var6, var7, lagstructur(#,#)).

Where, *var1, var2* and *var3* are the exogenous independent variables, var4 and var5 are the predetermined independent variables, and *var6, var7 and var8* are the endogenous independent variables.

The description of the syntax for these estimators (xtdpdsys) is similar to the xtabond (explained in paragraphs above). The only difference is the use of the comand xtdpdsys. At the level of methodology, the main difference between both estimators (xtabond and xtdpdsys) is the treatment used for the building of instrumental variables. The first one use only instrumental variables in differences and the second one use instrumental variables in levels.

4.3. xtabond2 Estimator (Instrumental Variables in Differences and Levels)

Stata has some estimators that use the instrumental variables in level and differences. (*xtdpdsys* and *xtdpd*). However, the estimator *xtabond2* has some advantages regarding the latter ones. This estimator allows excluding the lags of the dependent variables as regressors.

To perform the model using this estimator it is necessary to install the command in Stata. For doing that, it must be written as "*findit xtabond2*" in the command bar.

As it has been mentioned, *xtabond* only use as instrumental variables the lags in differences. This will reduce the number of instruments used in the regression. In addition, *xtabond2* uses the lags in levels, increasing the size of the matrix (system equation) and the number of instruments of the endogenous variable(s). Therefore, the first one, *xtabond*, is recommendable when the period is long, while the last one, *xtabond2*, is better for a panel with a short period of time, given that it incorporates the instruments in levels, reducing the loses of information.

xtabond2 can use instruments in differences and in levels. This information is incorporated in the model with the following expressions; instruments in difference and levels (*gmmstyle*), only differences (command eq(*diff*)) or only levels (*command eq(level)).*

To run the analysis on Stata with *xtabond2*, the instructions are divided into two parts; the first one identifies the variables that we are going to analyze, and the second one indicates how those variables are going to be incorporated into the model (endogenous, predetermined or exogenous). This second part also introduces the restrictions. Both parts of the equation are separated by a comma.

First, we will introduce the dependent variable with its lags and then, the independent variables. If we want to incorporate the dependent variable as a regressor, this must be specified between the dependent and independent variables using the syntax of *l.vardep*, for the first lag of the dependent variable, *l(#)*. This same structure is used for the specification of independent variables through their lags.

There are two ways for giving instructions to Stata in the treatment of the variables.

a. *gmmstyle* o *gmm*: for endogenous and predetermined variables.

b. *ivstyle* o *iv*: for exogenous variables

xtabond2 doesn't require the postestimation for Sargan and Hansen test (overidentification and for the serial autocorrelation of the error term), because these tests are reported directly.

In the following lines we describe the syntax in XTABOND 2 for the use of exogenous, predetermined and endogenous variables.

4.3.1. Models with Independent Exogenous Variables

xtabond2 vardep l.vardep var1 var2 var3, gmm (l.vardep, lag (# #)) iv (var1 var2 var3) robust twostep

Where,

var1 and *var2* are exogenous variables, and l.vardep is the lag of the dependent variable used as regressor, with its instrument restrictions *-lags (# #)*. This regressor *l.vardep* may be avoided in the estimation- In this case will also be avoided in the second part of the equation *-gmm (l.vardep, lag(# #))-*.

Robust is the instruction for working with heterocedasticity.

4.3.2. Models with Independent Variables as Predetermined

There are three alternatives for the specification of predetermined variables that report the same results:

1. *gmm(var 4 var5)*
2. *gmm(var 4 var5, lag(. .)*
3. *gmm(var4 var5, lag(1 .)*

xtabond2 vardep l.vardep var4 var5, gmm (l.vardep, lag (# #)) gmm (var4 var5) robust twostep

The last syntax has used the first option. This indicates that var4 and var5 are predetermined variables, and therefore, they are specified with the command gmm. Moreover, the exogenous variables will use the command iv.

4.3.3. Models with Endogenous Independent Variables

In the following example, the variables *var6* and *var7* are endogenous regressors. Note that the main differences with the syntax of the predetermined variables are found in the specification of the lags. As in the description above, there are three alternatives of specification:

1. *gmm(l.(var6 var7))*
2. *gmm(l.(var6 var7, lag(2 .))*
3. *gmm(l.(var6 var7, lag(1 .))*

In the following description we use the first alternative:

xtabond2 vardep l.vardep var6 var7, gmm (l.vardep, lag (# #)) gmm(l.var6 var7) robust twostep

The variables (*var6 and var7*) are considered as endogenous using 3 or more lags (lag(3 .), and the independent variable is also introduced as endogenous regressor: *l.vardep*.

The independent variables can be introduced using one or more lags. This is expressed in the first part of the equation. (e.g, if we want to use one lag, it should be expressed with the command *l.*, ex: *l.(var6))*. This means that the variable *var* will be analyzed using its first lag.

The syntax is as follows:

- Using the first lag in the independents variable

xtabond2 vardep l.vardep l.(var6 var7), gmm (l.vardep, lag (# #)) gmm(l.(var6 var7)) robust twostep

- Using the first and second lag in the independent variables

xtabond2 vardep l.vardep l(1/2).(var6 var7), gmm (l.vardep, lag (# #)) gmm(l.(var6 var7)) robust twostep

4.3.4. Combination in the Treatment of the Variables

In this section, we introduce different type of variables. The instruction for Stata would be as follow:

xtabond2 vardep l.vardep var1 var2 var3 var4 var5 l.(var6 var7), ///

gmm(l.vardep, lag (# #) ///iv(var1 var2 var3) ///

gmm(var4 var5)///

gmm(l.(var6 var7)) robust twostep ///

Where, *var1 var2 var3* are exogenous variables, *var4 var5* predetermined variables and *var6 var7* endogenous variables, using the first lag.

The following table (Table 2) shows a summary of model estimations alternatives to deal with endogeneity and the corresponding command in Stata.

Table 2: Alternatives to perform models with endogeneity.

Model type	**Performance step**	**Comand**
Models with exogenous independent variables	1 Model performance to apply Sargan test, with exogenous variables	xtabond vardep var1 var2 var3 varn, lags(#) twostep estat sargan
	2 Model performance with vce(robust) option and, exogenous variables	*xtabond vardep var1 var2 var3, varn, lags(#) vce(robust) twostep*
	3 Arellano and Bond test	estat abond
Models with predetermined independent variables	1 Model performance to apply Sargan test, with predetermined variables	*xtabond vardepend var1 var2 var3, lags(#)twostep pre (var4, va5, lagstructur(#,#)) estat sargan*
	2 Model performance with vce(robust) option (heterocedasticity) and predetermined variables	*xtabond vardepend var1 var2 var3, lags(#) twostep vce(robust) pre(var4, var5, lagstructur(#,#))*
	3 Arellano and Bond test	estat abond
Models with endogenous independent variables	1 Model performance to apply Sargan test, with predetermined endogenous variables	*xtabond vardepend var1 var2 var3, lags(#)twostep endog (var6, va7, lagstructur(#,#)) estat sargan*
	2 Model performance with vce(robust) option (heterocedasticity) and predetermined endogenous variables	*xtabond vardepend var1 var2 var3, lags(#) twostep vce(robust) endog(var6, var7, lagstructur(#,#))*
Endogenous models	Models with independent exogenous variables	*xtabond2 vardep l.vardep var1 var2 var3, gmm (l.vardep, lag (# #)) iv (var1 var2 var3) robust twostep*
	Models with independent variables as predetermined	*xtabond2 vardep l.vardep var4 var5, gmm (l.vardep, lag (# #)) gmm (var4 var5) robust twostep*
	Models with endogenous independent variables	*xtabond2 vardep l.vardep l.(var6 var7), gmm (l.vardep, lag (# #)) gmm(l.(var6 var7)) robust twostep*
	Models with mix of independent variables	*xtabond2 vardep l.vardep var1 var2 var3 var4 var5 l.(var6 var7), /// gmm(l.vardep, lag (# #) ///iv(var1 var2 var3) ///gmm(var4 var5)/// gmm(l.(var6 var7)) robust twostep ///*

4.4. Sargan, Hansen and Arellano and Bond Test. Interpretation

4.4.1. Sargan Test

This test verifies the validity of the instruments used in the analysis (Roodman, 2008). This test is used for One Step estimations and in samples where there is not a risk of overestimation. However, in Two Step estimations is recommended the use of the Hansen test (this last one is available for *xtabond2*) to check the *overidentification*.

The statistics reported is x^2. The number close to the x^2 in parentheses, correspond to the quantity of instruments over the instruments needed. The difference between the total instruments and the instruments leftover, is the optimal number of instrument for the model.

The interpretation of the Sargan test will be as follow:

Null hypothesis

Ho: All the restrictions of overidentification are valid.

Criteria of rejection or acceptation:

Prob$>chi^2 \geq 0.05(5\%)$

If the probability obtained is equal or higher to 0.05, the used instruments in the estimation are valid, and therefore overidentification doesn't exit. Therefore, there is no evidence to reject the *null hypothesis*. However, if the probability is lower than 0.05, the data is suggesting that the instruments are not valid and as consequence there is overidentification in the model. Therefore, we reject the null hypothesis.

If the probability is close to 1, this doesn't mean that the instruments are valid. It means that the asymptotic properties of the test have not been applied. In that case, we should reject Ho, as in the case where the probability is lower than < 0.05 (Roodman 2009).

Given that the estimator uses the higher quantity of available instruments and the probability of overindentification is high, when we reject the Sargan test, it is recommendable to apply some restrictions to the generation of instruments. For doing that, we can use the following commands:

xtabond and xtdpdsys: maxlags or maxldep

xtabond2: lags, collapse, eq(level) and eq(diff).

4.4.2. Hansen Test

This test is available for *xtabond2* and it is calculated directly when we use this command. In addition, it is recommendable to use it with the heterocedastic weight matrix (Two step). The interpretation of the test will be as follow:

Null hypothesis(Ídem Sargan)

Ho: All the restrictions of overidentification are valid.

Criteria of rejection/acceptation

$Prob > x^2 \geq 0.05(5\%)$

If the probability is close to 1, it means that the asymptotic properties of the test have not been applied, and therefore we also must reject Ho (Roodman 2009).

As recommendation $P(x^2)$ should be in the range of $0.05 \leq P(x^2) < 0.8$, being the optimal to find a probability $0.1 \leq (x^2) < 0.25$

If $P(X^2)$ is out of that range, the model could be overidentified and might be needed the introduction of some restrictions in the generation of instruments. For the application of this test Stata use the following commands:

Sargan Test: *estat sargan*

Using it after the estimation with One step

Hansen Test: it is given directly when xtabond2 is used.

4.4.3. Arellano and Bond Autocorrelation Test

Dynamic panel data introduces the condition of no correlation in the errors term (Cameron & Trivedi 2009). For testing that, we use the Arellano and Bond test.

We should expect that the probability of Ar(2) ($pr > z$) will be not significant at 5%. This will confirm the absence of serial autocorrelation in the errors. Normally, Ar(1) should be significant at 5% (AR (1) $pr > z < 0.05$).

The interpretation of this test will be as follow:

Null hypothesis:

Ho: Autocorrelation doesn't exit.

Criteria of rejection/acceptation

To reject that null hypothesis we will use AR (2). This rejection applies when the probability $pr > z$ is higher than 0.05, that is to say, the errors term are not serially correlated.

5. Conclusions

Panel data methodology has become one of the most popular tools used by researchers and academics who try to explain economics phenomenon by empirical analysis. Panel data allows incorporating into the analysis the effect of individuals and time, which gives a great advantage over cross sectional or time series.

Findings and new contributions have enabled to perform dynamic models, being able to analyze endogenous processes as evolutionary theory proposes. Most of the works in this regard have been conducted using databases made up of a large number of individuals and a short period of time (typical of micro data), however when estimating panels with few individuals and extended periods of time some limitations arise.

The main restriction in these cases arises from the overidentification of the model due to the proliferation of instruments of endogenous regressors when we use the GMM alternative including equations in levels and differences. This situation requires adjustments (options) and to apply several considerations to properly estimate models with this type of databases and methodology.

Although there are important advances in the study and works on panel data, dynamic version still requires additional efforts. Therefore, this article addresses this restriction in order to guide researcher to implement dynamic panel data using Stata software. In particular, this paper provides guidance for the scholars to understand the origin of the overidentification, as well as provide some tools to solve it.

Among the main strategies to conduct the overidentification described in this article are: restricting the lags of the dependent variable used as regressor of the model; limit the use of lags to generate instruments of endogenous independent variables; and avoid using equations in levels and differences simultaneously. In addition, researchers should pay attention to the serial autocorrelation tendency in this type of models. Thus, both challenges must be simultaneously addressed.

All the above must be permanently checked through the statistical tests in order to verify that conditions and restrictions of the estimation are found. Therefore, the incorporation of explanatory variables should be step by step, avoiding overidentification and allowing a better fit of the model.

This article is not free of weaknesses, since the objective is to provide a practical methodological support for non-specialist researchers in econometrics. The focus of this work is not the building of a theory, the search of a new estimator or specific test for this type of panel data, but this paper only tries to provide a guide to estimate dynamic models using Panel. In this sense, this work offers a way to carry out quantitative studies on several phenomena from data collected in a long time series and small number of individuals, which is common in database of countries, regions or where the observed unit has a limited population.

[Received: January 2017 — Accepted: September 2017]

References

Anderson, T. W. & Hsiao, C. (1981), 'Estimation of dynamic models with error components', *Journal of the American statistical Association* **76**(375), 598–606.

Arellano, M. & Bond, S. (1991), 'Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations', *The Review of Economic Studies* **58**(2), 277–297.

Arellano, M. & Bover, O. (1995), 'Another look at the instrumental variable estimation of error-components models', *Journal of Econometrics* **68**(1), 29–51.

Balestra, P. & Nerlove, M. (1966), 'Pooling cross section and time series data in the estimation of a dynamic model: The demand for natural gas', *Econometrica: Journal of the Econometric Society* **34**(3), 585–612.

Barro, R. (1991), 'Economic growth in a cross section of countries', *Quarterly Journal of Economics* (106), 407–443.

Blundell, R. & Bond, S. (1998), 'Initial conditions and moment restrictions in dynamic panel data models', *Journal of Econometrics* **87**(1), 115–143.

Cameron, A. & Trivedi, P. (2009), *Microeconometrics using Stata*, Stata Press College Station, Texas, United States.

Chudik, A., Pesaran, M. H. & Tosetti, E. (2011), 'Weak and strong cross-section dependence and estimation of large panels', *Structural Change and Economic Dynamics* **14**(1).

Dosi, G. (1988), 'Sources, procedures, and microeconomic effects of innovation', *Journal of economic literature* **XXVI**, 1120–1171.

Hoechile, D. (1933), 'Analysis of a complex of statistical variables into principal components', *Journal of Educational Psychology* (24), 417–520.

Labra, R. & Torrecillas, C. (2014), Guía cero para datos de panel. un enfoque práctico, Working paper 16, Universidad Autónoma de Madrid, Madrid, España.

Lee, K., Pesaran, M. & Pierse, R. (1990), 'Testing for aggregation bias in linear models', *Economic Journal* (100), 137–150.

Álvarez, I. & Labra, R. (2014), 'Technology Gap and Catching up in Economies Based on Natural Resources: The Case of Chile', *Journal of Economics, Business and Management* **3**(6), 619–627.

Maddala, G. (1971), 'The likelihood approach to pooling cross section and time series data', *Econometrica* **39**(6), 939–953.

Maddala, G. (1975), Some problems arising in pooling cross-section and time-series data, Discussion paper, University of Rochester, Nueva York.

Mairesse, J. & Griliches, Z. (1988), 'Heterogeneity in panel data: are there stable production functions?'.

Mileva, E. (2007), *Using Arellano-Bond Dynamic Panel GMM Estimators in Stata*, Fordham University, New York.

Nelson, R. & Winter, S. (1982), *An evolutionary theory of economic change*, Harvard University Press, United States.

Nerlove, M. (1971), 'Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross Sections', *Econometrica, Econometric Society* **39**(2), 359–382.

Pesaran, M. H. (2006), 'Estimation and inference in large heterogeneous panels with a multifactor error structure', *Econometrica* **74**(4), 967–1012.

Pesaran, M. H., Pierse, R. G. & Kumar, M. S. (1989), 'Econometric analysis of aggregation in the context of linear prediction models', *Econometrica: Journal of the Econometric Society* (57), 861–888.

Pesaran, M. H. & Smith, R. (1995), 'Estimating long-run relationships from dynamic heterogeneous panels', *Journal of econometrics* **68**(1), 79–113.

Phillips, P. C. B. & Sul, D. (2003), 'Dynamic Panel Estimation and Homogeneity Testing under Cross Section Dependence', *The Econometrics Journal* (6), 217–259.

Pérez-López, C. (2008), *Econometría Avanzada: Técnicas y Herramientas*, Pearson Prentice Hall, Madrid, España.

Roodman, D. (2006), *How to do xtabond2: An introduction to difference and system GMM in Stata*.

Roodman, D. (2009), 'A note on the theme of too many instruments', *Oxford Bulletin of Economics and Statistics* **71**(1), 135–158.

Ruíz-Porras, A. (2012), Econometric research with panel data: History, models and uses in mexico, MPRA-paper 42909, University Library of Munich, Germany.

Santos-Arteaga, F. J., Torrecillas, C. & Tavana, M. (2017), 'Dynamic effects of learning on the innovative outputs and productivity in spanish multinational enterprises', *The Journal of Technology Transfer* pp. 1–35.

Sargan, J. D. (1958), 'The estimation of economic relationships using instrumental variables', *Econometrica: Journal of the Econometric Society* **26**, 393–415.

Torrecillas, C., Fischer, B. B. & Sánchez, A. (2017), 'The dual role of R&D expenditures in European Union's member states: short-and long-term prospects', *Innovation: The European Journal of Social Science Research* **30**(4), 433–454.

Wooldridge, J. (2013), *Introductory Econometrics: A Modern Approach*, 5 edn, South-Western, Australia.