BBR BRAZILIAN BUSINESS REVIEW
v.10, n.1
Vitória-ES, jan.-mar. 2013
p. 127 - 149 ISSN 1808-2386 DOI: http://dx.doi.org/10.15728/bbr.2013.10.1.6

Panel data in accounting and finance: theory and application

Luiz Paulo Lopes Fávero†
University of São Paulo

ABSTRACT

The use of models that involve longitudinal data in accounting and finance is common. However, there is often a lack of proper care regarding the criteria for adopting one model over another as well as an insufficiently detailed discussion of the possible estimators to be studied in each situation. This article presents, in conceptual and applied form, the main panel data estimators that can be used in these areas of knowledge and discusses the definition of the most consistent model to be adopted in function of the data characteristics. The models covered for short panels are the POLS with clustered robust standard errors, with between estimator, fixed effects, fixed effects with clustered robust standard errors, random effects and random effects with clustered robust standard errors. In turn, for long panels, the models discussed are fixed effects, random effects, fixed effects with AR(1) error terms, random effects with AR(1) error terms, POLS with AR(1) errors and pooled FGLS with AR(1) errors. The models are also applied to a real case, based on data from Compustat Global. At the end, the main routines for applying each of the models in Stata are presented.

Keywords: Panel data; accounting; finance; estimation methods.

Received 05/02/2011; accepted 02/26/2012; published 03/26/2013.

*****Corresponding Author***:***

†. Professor Habilitation from the University of São Paulo.
Affiliation: Associate Professor at the University of São Paulo
Address: Av. Professor Luciano Gualberto, 908, FEA 3, sala 211- Cidade Universitária, São Paulo – SP – Brazil
E-mail: lpfavero@usp.br
Telephone: (11)3091-5820

Editor's Note: *This article was accepted by Bruno Funchal*

1 INTRODUCTION

The use of models that involve data originating from several *cross-sections* over time (panel data) in accounting and finances is increasingly growing and important. As much of the data from companies, cities or countries are released periodically, the researcher is invited, naturally, to apply longitudinal models to the study of phenomena that suffer the influence of the differences between individuals and its own temporal evolution.

According to Marques (2000), the main advantage of the use of models in data panels refers to the control of individual heterogeneity, that is, to the possibility of measuring separately the effects generated because of the existing differences between each observation in each cross-section, as well as being possible to evaluate the evolution, for a specific individual, of the variables in a study over time.

On the other hand, still according to Marques (2000), the panel data provide a larger quantity of information, greater variability of data, lesser collinearity between the variables, greater number of degrees of freedom and greater efficiency in the estimation. The inclusion of the dimension in cross-section, in a temporal study, confers a greater variability to the data, as the use of the aggregate data results in softer series that the individual series that function as a basis. This increase in data variability contributes to the reduction of an eventual collinearity existing between the variables.

According to Lima (2007), since Balestra and Nerlove (1966), several studies have been published with the use of panel data, with highlights for the works of Anderson and Hsiao (1982), Bhargava, Franzini and Narendranathan (1982), Bhargava and Sargan (1983), Baltagi and Griffin (1984), Holtz-Eakin, Newey and Rosen (1988), Breusch, Mizon and Schmidt (1989), Maddala (1993), Islam (1995), Ahn and Schmidt (1997), Krishnakumar and Ronchetti (2000), Hsiao (2003), Arellano (2003), Frees (2004), Baltagi (2008), Mátyás and Sevestre (2008), and Wooldridge (2010).

Marques (2000) made an important contribution, since, besides presenting the main concepts referring to the data panel models, he prepared a comparison between the different estimators used by the various authors in studies with micro and macroeconomic data.

In accounting and finances, on the other hand, the use of models that take into consideration the longitudinal aspects of accounting and financial phenomena are still incipient. In Brazil, we point out the works of Pimentel (2006), Souza (2006), Lima, Lima, Fávero and Galdi (2007), Coelho (2007), Nakamura, Martin, Forte, Carvalho Filho, Costa and Amaral (2007), Carneiro and Sherris (2008), Lima (2009), Bastos and Nakamura (2009), Almeida (2010), Malacrida, Lima, Fávero and Lima (2010) and Fávero and Sotelino (2011).

However, there is still, in this area, a lack of care as to the criteria for the adoption of one model in detriment of another, as well as the absence of a more detailed discussion about the possible estimators to be studied in each situation. In other words, the use of panel data in accounting and finances is, at times, elaborated without a deeper concern for the choice of a better model to be used, that is, little has been discussed about the adequacy of the use of the technique and about the definition of the best estimators. In this sense, the recent works of Pimentel (2009) and Jones, Kalmi and Mäkinen (2010) deserve to be highlighted.

The purpose of this article is to present, in a conceptual and structured manner, the main data estimators in a panel that can be used in accounting and finances, as well as help in the definition of the most consistent model to be adopted, as a function of the data characteristics. Also, this article has the objective of applying these models to a real case, based on data from *Compustat Global*. Finally, we present the main routines for the application of each of the models in Stata, since it believes that such procedures may provide a better relation between theory and practice, and to facilitate the implementation of the models in future research.

Thus, the present study does not have the intention to suggest the application of panel data in a given situation, as this depends fundamentally on the research issue and the data available to the researcher. The purpose, if this technique is to be used, is to assist in the correct application, with a view to determining the most appropriate models to reality and focused on decision-making.

Section 1 provides a conceptual review of the main panel data estimators and makes a distinction between models in short panel (with a larger number of individuals than the analysis period) and in long panel (with the larger number of periods than the number of individuals in the study). Section 2 presents an application of the main models presented and discusses the procedure of defining the best model through a review of the results. And finally, Section 3 presents the final considerations.

2 PANEL DATA MODELS

There are many different models that can be used for panel data. The basic distinction between them, according to Greene (2007), is the existence of fixed or random effects. The term "fixed effects" gives a wrong idea of modeling, because, in both cases, the effects at the individual level (firms, government agencies or countries, for example) are random. In this manner, according to Cameron & Trivedi (2009), the fixed effects models have the added complication that the regressors are correlated with the effects on the individual level and, therefore, a consistent estimation of the model parameters requires the elimination or control of the fixed effects. In this manner, a model that takes into account the specific effects of the individual for a dependent variable y_{it} specifies that:

$$y_{it} = \beta_{0i} + x'_{it}\beta_1 + \varepsilon_{it} \tag{1}$$

In which x_{it} are regressors, β_{0i} are the specific random effects for the individual and ε_{it} represents the idiosyncratic error.

With the error term being $\mu_{it} = \beta_{0i} + \varepsilon_{it}$ and x'_{it} correlated with the invariant error term in time (β_{0i}), it is supposed that x'_{it} is not correlated with the idiosyncratic error ε_{it}. The fixed effects model implies that $E(y_{it}|\beta_{0i}, x_{it}) = \beta_{0i} + x'_{it\beta1}$, assuming that $E(\varepsilon_{it}|\beta_{0i}, x_{it}) = 0$, so that $\beta_j = \partial E(y_{it}|\beta_{0i}, x_{it})/\partial x_{j,it}$. The advantage of the fixed effects model is that a consistent estimator of the marginal effect of the jnth regressor of $E(y_{it}|\beta_{0i}, x_{it})$ can be obtained, given that $x_{j,it}$ varies in time.

In the random effects model, on the other hand, it is assumed that β_{0i} is purely random, that is, it is not correlated to the regressors. The estimation, therefore, is prepared with an FGLS (*feasible generalized least squares*) estimator. The advantage of the random effects model is that it estimates all coefficients, even the time-invariant regressors, and, therefore, the marginal effects. Besides, $E(y_{it}|x_{it})$ can be estimated. But the major drawback is that these estimators are inconsistent if the fixed effects model is more appropriate.

As previously discussed, the dependent variable and the regressors can potentially vary simultaneously over both time and between individuals. While the variation over time or for any given individual, is known as *within* variance, the variation among individuals is called *between* variance. According to Wooldridge (2010), in the fixed effects model, the coefficient of a regressor with low variation *within* will be loosely estimated and will not be identified if not *within variance*. Therefore, it is of fundamental importance to differentiate between these variations to define the best model for panel data.

The total variation of the observations of a regressor x around the overall average $\bar{x} = 1/\sum_i T_i \sum_i \sum_t x_{it}$ in the data set can be decomposed in the sum of the *within* variation over time for each individual around $\bar{x}_i = 1/T\sum_t x_{it}$ and in the variation *between* individuals (for $\bar{x}_i$ around $\bar{x}$). According to Cameron & Trivedi (2009):

Variance *Within*: $s^2_{xW} = \frac{1}{\sum_i T_i - 1}\sum_i \sum_t (x_{it} - x_i + \bar{x})^2$

Variance *Between*: $s^2_{xB} = \frac{1}{N-1}\sum_i (\bar{x}_i - \bar{x})^2$

General Variance: $s^2_{xO} = \frac{1}{\sum_i T_i - 1}\sum_i \sum_t (x_{it} - \bar{x})^2$

Notations N and $\Sigma_i T_i$ correspond, respectively, to the number of individuals and the total number of observations over time. When submitting the application of panel data in this article, the variances of each of the regressors will be presented and discussed.

Also according to Cameron & Trivedi (2009), the β_1 parameter estimators of a fixed effects model to the equation (1) eliminate the β_{0i} fixed effects, that is, is prepare a transformation *within* by differentiation of averages. In this manner, a *within* estimation produces a modeling with different data about the average, and one cannot estimate a coefficient of a regressor without a variation over time. Thus, the β_{0i} fixed effects in equation (1) can be eliminated by subtracting the averages of each individual $\bar{y}_i = \bar{x}'_i \beta_1 + \bar{\varepsilon}_i$ in the corresponding model, resulting in the *within* model, or average model differences:

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i)'\beta_1 + (\varepsilon_{it} - \bar{\varepsilon}_i) \qquad (2)$$

In which $\bar{x}_i = T_i^{-1}\sum_{t=1}^{T_i} x_{it}$ and the *within* estimator is the OLS estimator (*ordinary least squares*) of this model. According to Cameron & Trivedi (2009), since β_{0i} was eliminated, the OLS estimator offers consistent estimates of β_1, even if β_{0i} is correlated with x_{it}, as is the case of the fixed effects model.

The *between* estimator uses only the variation between individuals (*cross-sections*) and is the OLS estimator of a regression of $\bar{y}_i$ as a function of $\bar{x}_i$, presented below (equation (3)). By taking into account only the *cross-section* variations in the data, the coefficient of any regressor which is invariant between individuals may not be identified.

$$\bar{y}_i = \beta_0 + \bar{x}'_i \beta_1 + (\beta_{0i} - \beta_0 + \bar{\varepsilon}_i) \qquad (3)$$

The consistency of this estimator demands that the error term $(\beta_{0i} - \beta_0 + \bar{\varepsilon}_i)$ not be correlated with x_{it}, which occurs when β_{0i} is a random effect, but not when it is a fixed effect. According to Hsiao (2003), this estimator is rarely used because the random effects estimators end up being more consistent.

The random effects estimator, on the other hand, is a FGLS estimator in equation (1). So the random effects model is the model of individual effects:

$$y_{it} = x'_{it}\beta_1 + (\beta_{0i} + \varepsilon_{it}) \tag{4}$$

With $\beta_{0i} \sim (\beta_0, \sigma_\alpha^2)$ and $\varepsilon_{it} \sim (0, \sigma_\varepsilon^2)$. In this manner, the error term $\mu_{it} = \beta_{0i} + \varepsilon_{it}$ is correlated over time t, for a given observation i, with correlation:

$$\text{corr}(\mu_{it}, \mu_{is}) = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\varepsilon^2), \text{ for all } s \neq t \tag{5}$$

The random effects estimator is the FGLS estimator of β_1 of equation (4) given the correlations of the errors in the equation (5).

According to Cameron & Trivedi (2009), in models with heteroscedastic and autocorrelated errors, the GLS estimator (*generalized least squares*) can be calculated as an OLS estimator in a model that has uncorrelated homoscedastic errors, obtained from (4) by an appropriate linear transformation. In the case of the random effects model of equation (4), this transformed model is given by equation (6).

$$(y_{it} - \theta_i \bar{y}_i) = (1 - \theta_i)\beta_0 + (x_{it} - \theta_i \bar{x}_i)'\beta_1 + \{(1 - \theta_i)\beta_{0i} + (\varepsilon_{it} - \theta_i \bar{\varepsilon}_i)\} \tag{6}$$

A FGLS estimator is obtained substituting θ_i, which is given for a consistent estimate indicated by:

$$\theta_i = 1 - \sqrt{\sigma_\varepsilon^2 / (T_i \sigma_\alpha^2 + \sigma_\varepsilon^2)} \tag{7}$$

The estimator of the random effects will be consistent and fully efficient if the random effects model is appropriate, but will be inconsistent if the fixed effects model is appropriate, since the correlation between x_{it} and β_{0i} results in a correlation between the regressors and the error term in equation (6). Likewise, also according to Cameron & Trivedi (2009), if there are no fixed effects, then the random effects estimator is consistent but inefficient and therefore an estimation with clustered robust standard errors should be obtained.

The expression of the estimate by feasible generalized least squares of a regression coefficient of model (1), assuming random effects, becomes equal to the same coefficient estimated in a fixed effects model (*within* estimation) if $\hat{\theta}_i = 1$.

2.1 Short Panel

If there are no fixed effects but the errors show a correlation inside the panel, then the random effects estimator will be consistent but inefficient and therefore an estimation with clustered robust standard errors should be obtained. In this manner, for a short panel, where T <N, an estimation with robust clustered standard errors can be obtained by considering the premise that the errors are independent among individuals and that $N \rightarrow \infty$, that is, that $E(\varepsilon_{it}, \varepsilon_{js}) = 0$ for $i \neq j$, that $E(\varepsilon_{it}, \varepsilon_{is})$ not be restricted and that ε_{it} be heteroscedastic.

According to Cameron and Trivedi (2009), the initial step for the implementation of a model with panel data is the application of a model POLS (*pooled ordinary least squares*), which supposes that the regressors are exogenous and that the error term is μ_{it}, instead of decomposition $\alpha_i + \varepsilon_{it}$. Therefore:

$$y_{it} = \beta_0 + x'_{it}\beta_1 + \mu_{it} \tag{8}$$

The parameters of this model are estimated by OLS, but the inference requires that there be control of correlation *within* error μ_{it} for a given individual, being prepared using robust standard errors with clustering at the individual level.

2.2 Long Panel

For long panel data, that is, with many periods for a relatively smaller number of subjects, the individual effects β_{0i} can be incorporated into x_{it} as *dummy* variables for each period according to the following model:

$$y_{it} = \beta_{0i} + \gamma_t + x'_{it}\beta_1 + \varepsilon_{it} \tag{9}$$

so that there are many time effects γ_t (monthly, quarterly or yearly effects, for example). A model *pooled*, for T > N, in which the regressors x_{it} include the intercept, the temporal effect and, possibly a vector of variable of an individual, can be written as:

$$y_{it} = x'_{it}\beta_1 + \mu_{it} \tag{10}$$

Since T is greater than N, it becomes necessary to specify a model that considers the presence of serial correlation error (Beck and Katz, 1995). So, for long panel data, the models *pooled* with estimation methods OLS (POLS) and FGLS become more appropriate, since they allow the use of an AR(1) model for μ_{it} over time where μ_{it} is heteroscedastic (Hoechle, 2007). So:

$$\mu_{it} = \rho_i \mu_{i,t-1} + \varepsilon_{it} \tag{11}$$

In which the terms ε_{it} are not correlated in time, but with correlation between individuals different from zero ($\text{corr}(\varepsilon_{it}, \varepsilon_{is}) = \sigma_{ts}$).

Alternatively the inclusion of one *dummy* variable vector for each period, it is estimated, finally, a model of individual effects with AR(1) error terms, which is a better model than that which considers i.i.d. error terms. So:

$$y_{it} = \beta_{0i} + x'_{it}\beta_1 + \mu_{it} \tag{12}$$

Soon, according to Cameron & Trivedi (2009), this model will potentially generate more efficient parameter estimates. In this case, given the estimate of $\hat{\rho}$ in equation (11), first, it eliminates the effect of the AR(1) error and, as a result, eliminates the individual effect by applying the difference in averages. So, the modeling can consider β_{0i} as a fixed or random effect.

Following the presentation of the panel data models, it is explained that this work will apply ten different types of modeling in order to provide a better understanding of the different types of estimators and their conditions of use, as well as present models for the study of behavior of the returns of the stock prices of companies listed on stock exchanges in Latin American countries, in a longitudinal perspective. Table 1 shows these ten different types of models. In the appendix, are routines for the application of each of these models in the Stata software.

Table 1: Panel Data Models to Be Estimated

Model	Description
POLS with Clustered Robust Standard-Errors	$y_{it} = \beta_0 + x'_{it}\beta_1 + \mu_{it}$ OLS estimation with correlation control *within* of error μ_{it} over time.
Model with *Between* Estimator	$\bar{y}_i = \beta_0 + \bar{x}_i\beta_1 + (\beta_{0i} - \beta_0 + \bar{\varepsilon}_i)$ The *between* estimator only uses the variation of *cross-sections* and is the OLS estimator of the regression of a function of. The consistency of this estimator demands that the error term not be correlated with x_{it}. $\bar{y}_i$ $\bar{x}_i$ $(\beta_{0i} - \beta_0 + \bar{\varepsilon}_i)$
Fixed Effects	$y_{it} = \beta_{0i} + x'_{it}\beta_1 + \varepsilon_{it}$ The β_{0i} parameters can be correlated with the x_{it} regressors, which allows a limited form of

	endogeneity. It is supposed that x_{it} is not correlated with the ε_{it} idiosyncratic error.
Fixed Effects with Clustered Robust Standard-Errors	$y_{it} = \beta_{0i} + x'_{it}\beta_1 + \varepsilon_{it}$ The β_{0i} terms can be correlated with the x_{it} regressors, which allows a limited form of endogeneity. It is supposed that the errors are independent between individuals and that ε_{it} is heteroscedastic.
Random Effects	$y_{it} = x'_{it}\beta_1 + (\beta_{0i} + \varepsilon_{it})$ The β_{0i} parameters and the ε_{it} idiosyncratic error terms are independent and identically distributed (i.i.d.). The random effects estimator is the FGLS of β_1, given that $corr(\mu_{it}, \mu_{is}) = \sigma^2_\alpha / (\sigma^2_\alpha + \sigma^2_\varepsilon)$.
Random Effects with Clustered Robust Standard-Errors	$y_{it} = x'_{it}\beta_1 + (\beta_{0i} + \varepsilon_{it})$ If there are no fixed effects, but the errors present correlation *within*, the random effects estimator is consistent, but inefficient. Therefore, clustered robust standard errors must be obtained.
Fixed Effects (AR)1 Error Terms	$y_{it} = \beta_{0i} + x'_{it}\beta_1 + \mu_{it}$ With $\mu_{it} = \rho_i\mu_{i,t-1} + \varepsilon_{it}$. β_{0i} is considered as a fixed effect.
Random Effects AR(1) Error Terms	$y_{it} = \beta_{0i} + x'_{it}\beta_1 + \mu_{it}$ With $\mu_{it} = \rho_i\mu_{i,t-1} + \varepsilon_{it}$. β_{0i} is considered as a random effect.
Pooled with OLS Estimation Method and AR(1) Error Terms	$y_{it} = \beta_{0i} + \gamma_t + x'_{it}\beta_1 + \varepsilon_{it}$ With $\mu_{it} = \rho_i\mu_{i,t-1} + \varepsilon_{it}$, wherein the ε_{it} are serially not correlated, but with correlation between individuals equal to $corr(\varepsilon_{it}, \varepsilon_{is}) = \sigma_{ts} \neq 0$.
Pooled with FGLS Estimation Method and AR(1) Error Terms	$y_{it} = \beta_{0i} + \gamma_t + x'_{it}\beta_1 + \varepsilon_{it}$ Similar to *pooled* model with OLS estimation method, but with FGLS estimator.

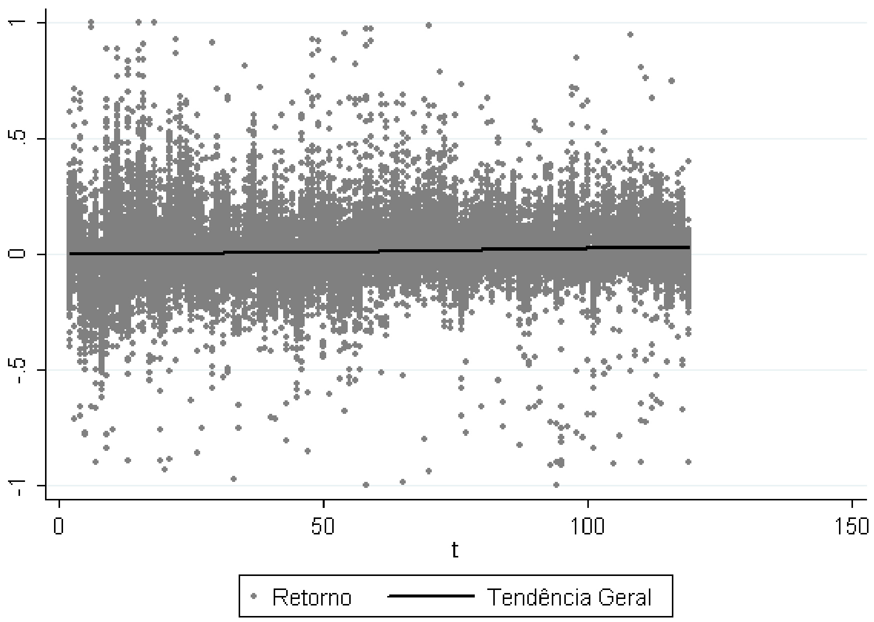
3 AN APPLICATION

After a discussion on the main data panel estimators, we present an application in financial accounting.

Since many of the accounting and financial data present a monthly, quarterly or annual release periodicity, it is common for studies in these areas using data models on short panels, since the number of individuals (companies, for example), exceeds the number of disclosure periods of data. On the other hand, nothing prevents the researcher from basing his or her study on a sample from companies in a given sector only, or use data with greater disclosure frequency (daily, for example), which could lead to a model with data on a long panel. Either way, it is essential that the identification of this database resource is done prior to the modeling itself. In this article, we will use two databases, being the first a short panel and the second a long panel.

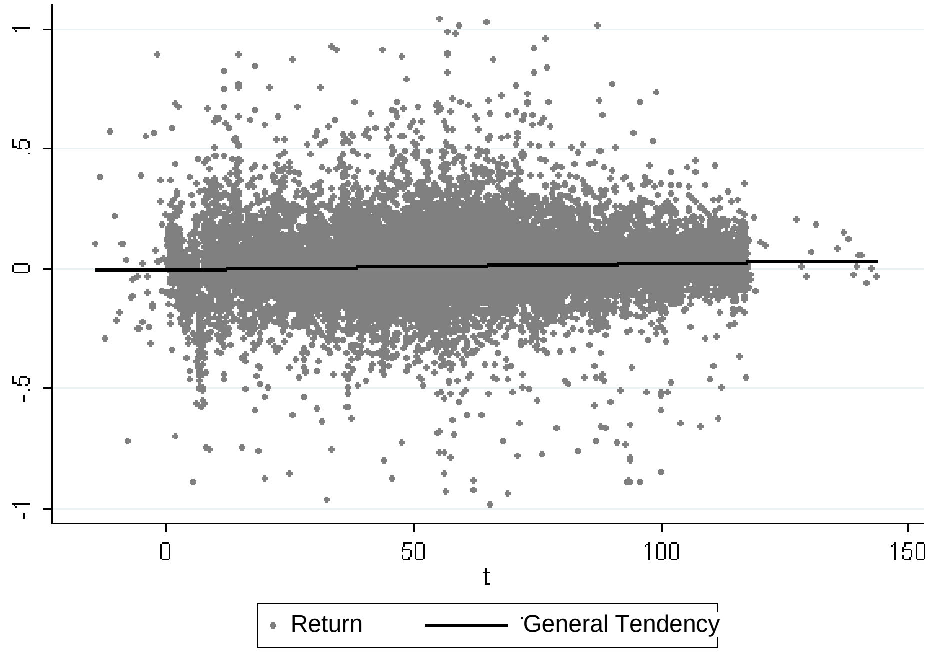
Initially, a *Compustat Global* base containing data on the profitability of shares from 473 companies from 7 countries in Latin America (Argentina, Brazil, Chile, Colombia, Mexico, Peru and Venezuela), for over 118 months (1998-2007), totaling 28,257 observations, will be used for the study of a short panel. Then, a stratum of the previous base, randomly selected, will be used, with data from only 40 companies, also during the 118-month period, totaling 4,720 observations, with a view to studying the long panel.

During the analyzed period, many companies showed significant growth rates in stock prices for one or more months. We considered, therefore, only monthly returns lower than 100%. Through graph 1 , we can see that these returns exhibit similar behaviors over time, although there are differences in the averages and slopes between individual series.

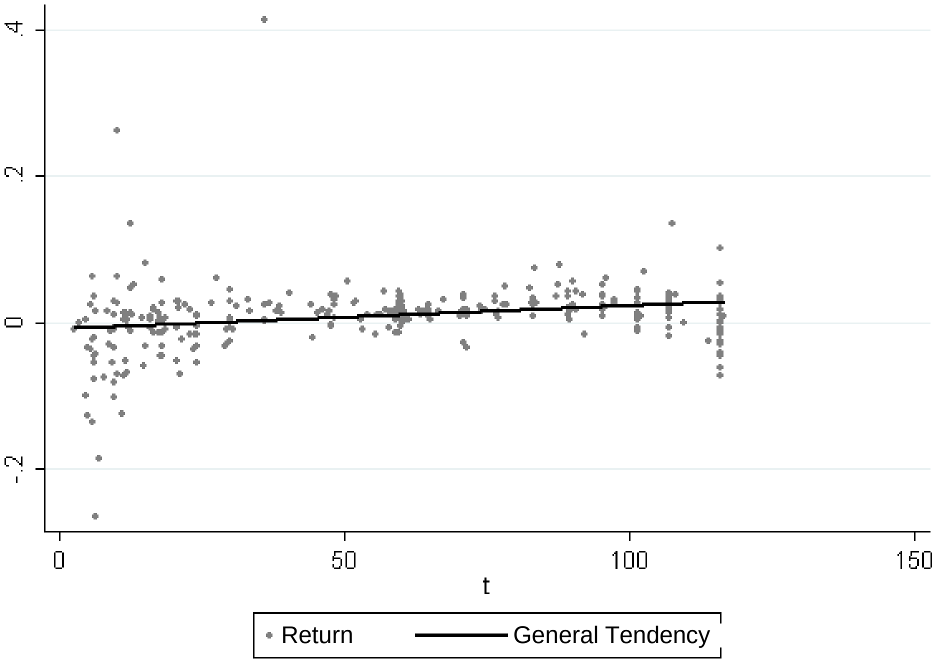


Graph 1: Evolution of Monthly Returns of Share Prices in Latin America

Each point on graph 1 represents one return pair of the stock price-month. This behavior suggests the preparation of longitudinal models, since the regressors may vary between companies and over time, as will be shown and discussed below. While graph 2 shows a variation of the monthly returns of the stock prices over time for each company, that is, shows the return deviations in relation to the individual average of each company (*within variation*), graph 3 presents the variation of the monthly returns between the companies, that is, showing the deviations of the monthly returns of the stock prices of the companies in relation to the general average for each instant of time (*between variation*).



Graph 2: Deviations of Monthly Returns in Relation to the Average of Each Company Over Time (*Within Variation*)



Graph 3: Deviation of Monthly Returns in Relation to the General Average for Every Moment of Time (*Between Variation*)

In this application, the use of the *Compustat Global* base is used in order to verify if the price-cash flow ratio is more significant than the price-earnings ratio per share to influence the monthly returns of share prices of companies in Latin America over time. According to Kennon (2010), as some investors prefer to use cash flow to make use of earnings per share for the evaluation of stock prices, since they argue that while the former is not easily manipulated, the same cannot be said for the second, this application provides an investigation on the subject, under a longitudinal perspective and with the use of several estimators.

As discussed, 10 different models of panel data will be developed with different considerations on the estimators and the error terms. The general model is given by:

$$\text{retorno}_{it} = \beta_{0i} + \beta_1.(\text{pcf})_{it} + \beta_2.(\text{pe})_{it} + \varepsilon_{it} \tag{13}$$

In which β_1 and β_2 represent the changes in the return of share prices when a unit of cash flow ratio (pcf) or price-to-earnings ratio per share (pe) occurs, respectively, *ceteris paribus*.

Below, we discuss the results of modeling, both for a short panel, as for a long panel.

Data Models for Short Panel

As the sample, in this case, it provides data from 473 companies in 118 months, the panel can be considered short (T <N) ..

Table 2 presents the variance decomposition for each of the regressors.

Table 2: Statistics of Short Panel and Variance Decomposition

Variable	Decomposition	Average	Standard Deviation	Minimum	Maximum	Comments
id (company)	general *between* *within*		 0.000			N.T = 28,257 N = 473
t (month)	general *between* *within*		35.492 33.315 26.489	1.000 1.500 -5.398	118.000 118.000 145.188	N.T = 28,257 N = 473
var_y (return)	general *between* *within*	0.013	0.141 0.036 0.139	-0.999 -0.134 -0.986	1.000 0.414 1.027	N.T = 28,257 N = 473
var_x1 (pcf)	general *between* *within*	7.246	246.040 212.442 221.678	-19,841.38 -2,168.16 -17,665.97	10,594.7 3,363.81 7,238.13	N.T = 20,224 N = 361
var_x2 (pe)	general *between* *within*	12.965	212.550 228.033 176.068	-5,790.53 -2,872.58 -4,521.99	12,251.82 3,889.48 8,375.30	N.T = 27,816 N = 463

According to table 1, note that the stock is time invariant and, therefore, presents the *within* variation equal to zero. ON the other hand, the variable referring time (month) is not invariant among companies, since this is an unbalanced panel and hence its *between* variation, even though lower than *within*, is not equal to zero. Of the remaining variables, only *pcf* presents a greater variation between individuals (*between*) than over time (*within*), but it is still not possible to declare that the *within* estimation will result in a loss of efficiency, since the proportion between the *within* and *between* variances of each variable is different and the statistical significances of each of these models are not yet known. Table 1, however, provides a greater basis for the adoption of models of panel data and the application of different estimators. The columns "Minimum" and "Maximum" show, respectively, the minimum and maximum values of x_{it} for the "general" line, $\bar{x}_i$ for the "*between*" line and $(x_{it} - \bar{x}_i + \bar{x})$ for the "*within*" line.

Table 3 presents the results of models considering 6 different estimators.

Table 3: Data Models for Short Panel

Variable	**POLS with Clustered Robust Standard-Errors**	***Between*** Estimator	**Fixed Effects**	**Fixed Effects with Clustered Robust Standard Errors**	**Random Effects**	**Random Effects with Clustered Robust Standard-Errors**
pcf	1.52×10^{-5}* (5.93×10^{-6})	1.24×10^{-5} (1.32×10^{-5})	1.52×10^{-5}* (4.96×10^{-6})	2.15×10^{-5}* (5.74×10^{-6})	1.68×10^{-5}* (4.62×10^{-6})	168×10^{-5}* (5.19×10^{-6})
pe	-1.09×10^{-5} (6.19×10^{-6})	-5.59×10^{-7} (1.1×10^{-5})	-1.28×10^{-5} (6.05×10^{-6})	-1.28×10^{-5} (5.16×10^{-6})	-1.16×10^{-5} (5.13×10^{-6})	-1.16×10^{-5} (4.54×10^{-6})
constant	0.013* (1.04×10^{-3})	0.009* (2.11×10^{-3})	0.013* (1.03×10^{-3})	0.013* (5.23×10^{-5})	0.011* (1.59×10^{-3})	0.011* (1.17×10^{-3})
N.T	20,224	20,224	20,224	20,224	20,224	20,224
R^2	6.0×10^{-4}					
R^2 (general)		4.0×10^{-4}	6.0×10^{-4}	6.0×10^{-4}	6.0×10^{-4}	6.0×10^{-4}
R^2 *(between)*		3.1×10^{-4}	1.0×10^{-3}	1.0×10^{-3}	1.0×10^{-3}	1.0×10^{-3}
R^2 *(within)*		8.0×10^{-4}	2.1×10^{-3}	2.1×10^{-3}	1.5×10^{-3}	1.5×10^{-3}
F	5.12	0.73	9.57	7.10		
sig. F	0.006	0.483	0.000	0.001		
Wald χ^2					14.06	10.83
sig. χ^2					0.001	0.004

Note: Standard Errors between brackets.
*sig. < 0.05.

As can be seen, the estimated coefficients vary from model to model, which reflects the existence of different results if the *within* or *between* variations can be used.

At first, one verifies, as to the appropriateness of the models, that the regressor's vector presents a statistical significance in all cases, except the model with the *between* estimator (sig. F for the POLS models, *between* and with fixed effects and sig. Wald χ^2 for models with random effects). As the R^2 statistics are considerably lower, these models are not relatively adequate for prediction purposes. An important result, however, relates to the existence of higher values for R^2 *within* in all models in which this statistic is calculated. Taking as basis the expressions of each of the statistics R^2

R^2 general: $\rho^2\{(y_{it} - \bar{y}_i), (x'_{it}\hat{\beta} - \bar{x}'_i\hat{\beta})\}$

R^2 *between*: $\rho^2(\bar{y}_i, \bar{x}'_i\hat{\beta})$

R^2 *within*: $\rho^2(y_{it}, x'_{it}\hat{\beta})$

In which $\rho^2(x,y)$ represents the square correlation between x and y. Note that the *within* estimators better explain the *within* variation in all models, even those with random effects.

Note also that the *pe* variable is not statistically significant (sig. > 0,05) in models presented in the presence of the *pcf* variable. The latter, with the exception of model with the *between* estimator, is significant in explaining the behavior of the returns of stock prices (sig. < 0,05), confirming the argument of some analysts in favor of using this variable.

As to the *pcf* variable, it appears that the standard errors in fixed effects and random effects models with clustered robust standard errors are larger than the respective models without this consideration. The regressors estimated in POLS models and *between* offer even greater standard errors, even with the *pcf* variable being statistically significant (sig. < 0,05) in the POLS model.

The Breusch-Pagan LM Test, applied after the modeling of random effects, helps in the rejection of the null hypothesis that there is adaptation in the POLS model in relation to the random effects model, since $\chi^2 = 70.7$ (sig. $\chi^2 = 0.000$). Following through, by means of the Chow F test, the null hypothesis that there is equality of intercepts and slopes for all companies (POLS) is rejected. Therefore, these parameters differ from those obtained by means of fixed effects models, since F = 2.34 (sig. F = 0.000). Finally, the Hausman test for fixed effects assists in rejecting the null hypothesis that the random effects model provides more consistent parameter estimates, since, for this case, $\chi^2 = 17.07$ (sig. $\chi^2 = 0.000$).

According to Islam (1995), the main use of panel data modeling is its ability to allow differences occur between countries, which means that the results are significantly different from those obtained by isolated regressions for each country. In table 4, the regression coefficients are presented for each of the sample countries.

Table 4: Coefficients by Country

Country	pcf	constant
Argentina	-1.97×10^{-5}	0.006
Brazil	-1.63×10^{-5}	0.018
Chile	2.08×10^{-5}	0.007
Colombia	9.07×10^{-6}	0.015
Mexico	-2.66×10^{-5}	0.007
Peru	8.75×10^{-5}	0.016
Venezuela	1.32×10^{-4}	0.010

Note: Dependent Variable: Monthly Return of the Stock Price.

Although the price-cash flow ratio is more significant in explaining the returns of stock prices in the countries of Latin America, Table 4 reveals the existence of different influences. The different coefficients and signs of the *pcf* variable and the constant express the importance of considering the modeling panel data and provide the formulation of future research on the economic reasons why countries have different behaviors in the prices of actions of their companies over time.

Data Models for Long Panel

For this case, as the sample provides data from 40 companies over 118 months, the panel can be considered long (T> N). Thus, the influence time is very important in long series, models of random and fixed effects are also applied with the consideration of autoregressive (AR(1)) components for the waste, which can result in more efficient parameter estimates for long panels.

As was prepared for the short panel, table 5 shows the variance decomposition for each of the regressors of the long panel.

Table 5: Statistics of Long Panel and Variance Decomposition

Variable	Decomposition	Average	Standard Deviation	Minimum	Maximum	Comments
id (company)	general *between* *within*		 0.000			N.T = 4,720 N = 40
t (month)	general *between* *within*		34.066 0.000 34.066	1.000 60.500 1.000	118.000 60.500 118.000	N.T = 4,720 N = 40
var_y (return)	general *between* *within*	0.012	0.144 0.005 0.144	-0.662 0.005 -0.656	0.899 0.022 0.896	N.T = 4,720 N = 40
var_x1 (pcf)	general *between* *within*	5.312	22.261 5.106 21.682	-264.68 -3.262 -270.274	107.37 10.906 101.776	N.T = 4,720 N = 40
var_x2 (pe)	general *between* *within*	14.488	66.362 12.089 65.279	-536.13 1.365 -537.894	735.86 40.972 734.095	N.T = 4,720 N = 40

Here 9 companies were purposely chosen so that the variable relating to time (month) would be invariant, meaning that the panel would be balanced, so that its *between* variation would be equal to zero. All other variables showed less variation between individuals (*between*) than over time (*within*), but it is also not possible to say that the *between* estimation will result in a loss of efficiency.

In the same way as was done for the short panel, table 6 shows the results of the models, considering also six different estimators.

Table 6: Data Models for Long Panel

Variable	**Fixed Effects**	**Random Effects**	**Fixed Effects with AR(1) Errors**	**Fixed Effects with AR(1) Errors**	***Pooled* OLS with AR(1) Errors**	***Pooled* FGLS with AR(1) Errors**
pcf	-5.68×10^{-5} (9.74×10^{-5})	-7.55×10^{-5} (9.45×10^{-5})	-5.77×10^{-5} (4.32×10^{-5})	-6.49×10^{-5} (4.88×10^{-5})	-7.13×10^{-5} (2.36×10^{-4})	-1.49×10^{-4} (1.97×10^{-4})
pe	-8.42×10^{-5}* (3.24×10^{-5})	-9.11×10^{-5}* (3.17×10^{-5})	-7.88×10^{-5}* (1.70×10^{-5})	-8.58×10^{-5}* (2.32×10^{-5})	-8.90×10^{-5} (6.90×10^{-5})	-8.63×10^{-5} (6.07×10^{-5})
constant	0.014* (2.23×10^{-3})	0.014* (2.22×10^{-3})	0.013* (2.20×10^{-3})	0.014* (2.21×10^{-3})	0.014 (8.22×10^{-3})	0,013 (7.24×10^{-3})
N.T	4,720	4,720	4,680	4,720	4,720	4,720
R^2 R^2 (general) R^2 *(between)* R^2 *(within)*	 1.9×10^{-3} 0.494 1.5×10^{-3}	 1.9×10^{-3} 0.504 1.5×10^{-3}	 1.9×10^{-3} 0.462 1.2×10^{-3}	 1.9×10^{-3} 0.499 1.5×10^{-3}	1.8×10^{-3}	
F sig. F	3.59 0.027		2.88 0.050			
Wald χ^2 sig. χ^2		9.00 0.011		7.87 0.048	1.72 0.422	2.57 0.276

Note: Standard Errors between brackets.
*sig. < 0.05.

According to table 5, we can see that the estimated coefficients also vary between models. First, note the existence of much higher standard errors in the fixed and random effects models (over 100%) compared with those reported by their respective models with AR(1) effects in the error terms. This fact may have occurred because of the nature of the panel under review, that is, by being long.

But even allowing that the error terms are correlated between companies, it is noted that there wasn't, in this case, a reduction of standard errors of the *pooled* models with OLS and FGLS estimators compared with those obtained previously by means of models of fixed and random effects with AR(1) error terms.

With respect to the suitability of the models themselves, there is the statistical significance of the set of variables in cases in which were considered fixed or random effects with or without AR(1) error terms. As presented during the preparation of models for the short panel, although there is relative importance of R^2 statistics for prediction effects, their values are not significantly elevated in the models under review.

The models of random and fixed effects offer an alternative for long panel data, wherein the individual effects are considered with AR(1) error terms, and represent the best models than those that consider the error terms i.i.d., which can generate more efficient parameter estimates. In fact, models with fixed and random effects with AR(1) error terms show standard errors on the order of 30% to 50% lower than those obtained by the respective models without consideration of AR(1) error terms.

The Hausman test applied to the fixed and random effects models with AR(1) error terms assists in rejecting the null hypothesis that the random effects model provides more consistent parameter estimates, since, in this case, $\chi^2 = 10.50$ (sig. $\chi^2 = 0.005$).

Finally, it is worth mentioning that, in this case, the results appear to be contrary to those obtained for the short panel, that is, the pcf variable is not statistically significant (sig. > 0.05) in the presence of the *pe* variable, which presents itself with a negative sign. However, as the companies considered in this case are originated only from Argentina, Brazil and Mexico, a more detailed investigation on the economic reasons behind this phenomenon needs to be performed. As the negative signs of the parameters of the regressors are consistent with those presented in table 3 for these countries, it emphasizes even more the importance of the correct application of the panel models for the study of existing differences between individuals and, over time, for a given phenomenon.

4 CONCLUSIONS AND RECOMMENDATIONS

Panel data models allow the researcher to evaluate the relationship between some performance variable and several predictor variables, allowing you to draw inferences about the possible differences between individuals and, over time, about the evolution of that which is to be studied. Given its characteristics, it is natural that much research in accounting and finance will make use of such models, since a good quantity of data is published with specific periodicity for companies, cities, states or countries.

To this end, it is necessary, as well as for any other econometric technique, that the application is accompanied by methodological rigor and certain precautions when analyzing the results, especially if they are aimed at forecasting. The adoption of a given estimator, in detriment of another that is considered biased or inconsistent, can aid the researcher in choosing the best model, enhancing its research and providing new studies on the chosen theme.

In this article, we sought to establish six different models for a specific short panel, and another six for a long panel. The data, in both cases, originated from *Compustat Global*. According to Makino, Isobe and Chan (2004), previous studies have used longitudinal modeling to study the variability of performance variables between companies or countries over time, with empirical research conducted in several areas of knowledge. But it is not common to find studies that are applied to stock markets, considering the differences between emerging economies.

The analysis of the contribution of price-cash flow ratio and price-to-earnings ratio per share on the monthly returns of stocks from emerging markets enables it to enhance the discussion on how the behaviors of markets in the emerging countries differ. But this approach was adopted only as an example within a larger goal, which was to present how different estimators can produce inconsistent results when producing panel data models, and help the researcher to choose the most suitable model both in the case of a short panel, as in the long panel.

The present article intends to contribute only with one part of the innumerable research projects that can arise. Models whose performance variables are presented in the form of the *dummy*, with censored data or count data present consistently different estimators and therefore specific routines in software like Stata. These models were not discussed in this article. It is expected, therefore, that this discussion is just the beginning and will gain importance in accounting and finances, given the vastness of research possibilities.

REFERENCES

AHN, S. C.; SCHMIDT, P. Efficient estimation of dynamic panel data models: alternative assumptions and simplified estimation. **Journal of Econometrics**, v. 76, n. 1-2, p. 309-321, 1997.

ALMEIDA, J. E. F. **Qualidade da informação contábil em ambientes competitivos.** 2010. 174 f. Tese (Doutorado em Controladoria e Contabilidade) – Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo.

ANDERSON, T. W.; HSIAO, C. Formulation and estimation of dynamic models using panel data. **Journal of Econometrics**, v. 18, n. 1, p. 47-82, 1982.

ARELLANO, M. **Panel data econometrics:** advanced texts in econometrics. New York: Oxford University Press, 2003.

BALESTRA, P.; NERLOVE, M. Pooling cross section and time series data in the estimation of a dynamic model: the demand for natural gas. **Econometrica**, v. 34, n. 3, p. 585-612, 1966.

BALTAGI, B. H. **Econometric analysis of panel data.** 4. ed. New York: John Wiley and Sons, 2008.

BALTAGI, B. H.; GRIFFIN, J. M. Short and long run effects in pooled models. I**nternational Economic Review**, v. 25, n. 3, p. 631-645, 1984.

BASTOS, D. B.; NAKAMURA, W. T. Determinantes da estrutura de capital das companhias abertas no Brasil, México e Chile no período 2001-2006. **Revista Contabilidade e Finanças**, v. 20, n. 50, p. 75-94, 2009.

BECK, N.; KATZ, J. N. What to do (and not to do) with time-series cross-section data. **American Political Science Review**, v. 89, n. 3, p. 634-647, 1995.

BHARGAVA, A.; FRANZINI, L.; NARENDRANATHAN, W. Serial correlation and the fixed effects model. **Review of Economic Studies**, v. 49, n. 4, p. 533-549, 1982.

BHARGAVA, A.; SARGAN, J. D. Estimating dynamic random effects models from panel data covering short time periods. **Econometrica**, v. 51, n. 6, p. 1635-1659, 1983.

BREUSCH, T. S.; MIZON, G. E.; SCHMIDT, P. Efficient estimation using panel data. **Econometrica**, v. 57, n. 3, p. 695-700, 1989.

CAMERON, A. C.; TRIVEDI, P. K. **Microeconometrics using Stata.** College Station: Stata Press, 2009.

CARNEIRO, L. A. F.; SHERRIS, M. Corporate interest rate risk management with derivatives in Australia: empirical results. **Revista Contabilidade e Finanças**, v. 19, n. 46, p. 86-107, 2008.

COELHO, A. C. D. **Qualidade informacional e conservadorismo nos resultados contábeis no Brasil.** 2007. 240 f. Tese (Doutorado em Controladoria e Contabilidade) – Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo.

FÁVERO, L. P. L.; SOTELINO, F. B. Elasticities of stock prices in emerging markets. In: BATTEN, J. A.; SZILAGYI, P. G (Ed.) The impact of the global financial crisis on emerging financial markets**.** Emerald Group Publishing Limited, 2011. **(Contemporary Studies in Economic and Financial Analysis, Vol. 93,** p. 473-493).

FREES, E. W. **Longitudinal and panel data:** analysis and applications in the social sciences. Cambridge: Cambridge University Press, 2004.

GREENE, W. H. **Econometric analysis.** 6. ed. Upper Saddle River: Prentice Hall, 2007.

HOECHLE, D. Robust standard errors for panel regressions with cross-sectional dependence. **Stata Journal**, v. 7, n. 3, p. 281-312, 2007.

HOLTZ-EAKIN, D.; NEWEY, W.; ROSEN, H. S. Estimating vector auto regressions with panel data. **Econometrica**, v. 56, n. 6, p. 1371-1395, 1988.

HSIAO, C. **Analysis of panel data.** 2. ed. Cambridge: Cambridge University Press, 2003.

ISLAM, N. Growth empirics: a panel data approach. **The Quarterly Journal of Economics**, v. 110, n. 4, p. 1127-1170, 1995.

JONES, D. C.; KALMI, P.; MÄKINEN, M. The productivity effects of stock option schemes: evidence from Finnish panel data. **Journal of Productivity Analysis**, v. 33, n. 1, p. 67-80, 2010.

KENNON, J. **Price to cash flow ratio**. Disponível em: <http://beginnersinvest.about.com/od/financialratio/a/pricetocashflow.htm>. Acesso em: 07 ago. 2010.

KRISHNAKUMAR, J.; RONCHETTI, E. (Ed.) **Panel data econometrics**: future directions. Amsterdam: North Holland, 2000.

LIMA, G. A. S. F. Nível de evidenciação x custo da dívida das empresas brasileiras. **Revista Contabilidade e Finanças**, v. 20, n. 49, p. 95-108, 2009.

LIMA, G. A. S. F. **Utilização da teoria da divulgação para avaliação da relação do nível de *disclosure* com o custo da dívida das empresas brasileiras**. 2007. 98 f. Tese (Doutorado em Controladoria e Contabilidade) - Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo.

LIMA, G. A. S. F. et al. Influência do *disclosure* voluntário no custo de capital de terceiros. In: CONGRESSO USP DE CONTROLADORIA E CONTABILIDADE, 7., 2007, São Paulo. **Anais**... Universidade de São Paulo, 2007.

MADDALA, G. S. **The econometrics for panel data.** Brookfield: Elgar, 1993.

MAKINO, S.; ISOBE, T.; CHAN, C. M. Does country matter? **Strategic Management Journal**, v. 25, n. 10, p. 1027-1043, 2004.

MALACRIDA, M. J. C. et al. An accounting accruals model to predict future operating cash flows: evidence from Brazil. **International Journal of Management**, v. 27, n. 3, p. 562-578, 2010.

MARQUES, L. D. **Modelos dinâmicos com dados em painel:** revisão da literatura. Série *Working Papers* do Centro de Estudos Macroeconômicos e Previsão (CEMPRE) da Faculdade de Economia do Porto, Portugal, n. 100, 2000.

MÁTYÁS, L.; SEVESTRE, P. (Ed.). **The econometrics of panel data**: fundamentals and recent developments in theory and practice. 3. ed. New York: Springer, 2008.

NAKAMURA, W. T. et al. Determinantes de estrutura de capital no mercado brasileiro: análise de regressão com painel de dados no período 1999-2003. **Revista Contabilidade e Finanças**, v. 18, n. 44, p. 72-85, 2007.

PIMENTEL, R. C. **Accounting earnings properties and determinants of earnings response coefficient in Brazil**. 2009. 162 f. Tese (Doutorado em Controladoria e Contabilidade) – Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo.

PIMENTEL, R. C. **O mercado de *eurobonds* e as captações brasileiras**: uma abordagem empírico-descritiva. 2006. 178 f. Dissertação (Mestrado em Controladoria e Contabilidade) – Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo.

SOUZA, M. S. **Fluxo de caixa por regime de competência.** 2006. 93 f. Dissertação (Mestrado em Controladoria e Contabilidade) – Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo.

WOOLDRIDGE, J. M. **Econometric analysis of cross section and panel data**. 2. ed. Cambridge: MIT Press, 2010.

APPENDIX: ROUTINES IN STATA

Definition of the Panel:
```
xtset id t
```

Preparation of Graph 1:
```
graph twoway (scatter var_y t) (lfit var_y t)
```

Preparation of Graph 2:
```
preserve
xtdata, fe
graph twoway (scatter var_y t) (lfit var_y t)
restore
```

Preparation of Graph 3:
```
preserve
xtdata, be
graph twoway (scatter var_y t) (lfit var_y t)
restore
```

Preparation of Panel Variance Decomposition Tables (Tables 1 and 4):
```
xtsum id t var_y var_x1 var_x2
```

Preparation of Data Models in Short Panel:

- POLS with Clustered Robust Standard-Errors
```
regress var_y var_x1 var_x2, vce(cluster id)
```
- Model with *Between* Estimator:
```
xtreg var_y var_x1 var_x2, be
```
- Fixed Effects:
```
xtreg var_y var_x1 var_x2, fe
```
- Fixed Effects with Clustered Robust Standard Errors:
```
xtreg var_y var_x1 var_x2, fe vce(cluster id)
```
- Random Effects:
```
xtreg var_y var_x1 var_x2, re
```
- Random Effects with Clustered Robust Standard Errors:
```
xtreg var_y var_x1 var_x2, re vce(cluster id)
```

Preparation of the Estimators Comparison Table for Short Panel Models (Table 2):
```
quietly regress var_y var_x1 var_x2, vce(cluster id)
estimates store POLS_rob
quietly xtreg var_y var_x1 var_x2, be
estimates store BE
quietly xtreg var_y var_x1 var_x2, fe
estimates store FE
quietly xtreg var_y var_x1 var_x2, fe vce(cluster id)
estimates store FE_rob
quietly xtreg var_y var_x1 var_x2, re
estimates store RE
quietly xtreg var_y var_x1 var_x2, re vce(cluster id)
estimates store RE_rob
```

```
estimates table POLS_rob BE FE FE_rob RE RE_rob, b se stats(N r2 r2_o r2_b r2_w F chi2) b(%7.5f)
```

Preparation of Breusch-Pagan LM Test:

```
xttest0
```

Preparation of Hausman Test for Short Panel:

```
hausman FE RE, sigmamore
```

Preparation of Coefficients Table for each Country (Table 3):

```
preserve
statsby, by(pais) clear: xtreg var_y var_x1 var_x2, fe
list, clean
restore
```

Preparation of Data Models in Long Panel:

- Fixed Effects:

```
xtreg var_y var_x1 var_x2, fe
```

- Random Effects:

```
xtreg var_y var_x1 var_x2, re
```

- Fixed Effects with AR(1) Errors:

```
xtregar var_y var_x1 var_x2, fe
```

- Random Effects with AR(1) Errors:

```
xtregar var_y var_x1 var_x2, re
```

- POLS with AR(1) Errors and correlation between individuals:

```
xtpcse var_y var_x1 var_x2, corr(ar1)
```

- FGLS with AR(1) Errors and correlation between individuals:

```
xtgls var_y var_x1 var_x2, corr(ar1) panels(correlated)
```

Preparation of the Estimators Comparison Table for Long Panel Models (Table 5):

```
quietly xtreg var_y var_x1 var_x2, fe
estimates store FE
quietly xtreg var_y var_x1 var_x2, re
estimates store RE
quietly xtregar var_y var_x1 var_x2, fe
estimates store FEAR1
quietly xtregar var_y var_x1 var_x2, re
estimates store REAR1
quietly xtpcse var_y var_x1 var_x2, corr(ar1)
estimates store POLSAR1
quietly xtgls var_y var_x1 var_x2, corr(ar1) panels(correlated)
estimates store FGLSAR1
estimates table FE RE FEAR1 REAR1 POLSAR1 FGLSAR1, b se stats(N r2 r2_o r2_b r2_w F chi2) b(%7.5f)
```

Preparation of Hausman Test for Long Panel:

```
hausman FEAR1 REAR1
```