

Endogeneity in panel data regressions: methodological guidance for corporate finance researchers

Lucas A. B. C. Barros¹
Daniel Reed Bergmann¹

¹University of São Paulo, School of Economics, Management, Accounting and Actuarial Science, São Paulo, Brazil

F. Henrique Castro²

²Fundação Getulio Vargas, São Paulo School of Economics, São Paulo, Brazil

Alexandre Di Miceli da Silveira³

³Alvares Penteado School of Business (FECAP), São Paulo, Brazil

Received on

09/10/2018

Approved on

06/28/2019

Responsible editor:

Prof. Dr. Joelson Oliveira
Sampaio

Evaluation process:

Double Blind Review

Abstract

Purpose – To describe the use of specific lags (and/or temporal differences) of the original regressors as instrumental variables in a succinct and practical way, showing, by means of a theoretical discussion illustrated by an original simulation exercise, how combining these with adequate modeling of firm and time fixed effects can address not only the dynamic endogeneity problem, but also those derived from the presence of omitted variables, measurement errors, and simultaneity between dependent and independent variables.

Design/methodology/approach – Monte Carlo simulation

Findings – The traditional OLS, RE, and FE estimators may be inconsistent in the presence of endogeneity problems that are quite plausible in the context of corporate finance. On the other hand, the estimation methods for panel data based on GMM that use assumptions of sequential exogeneity of the regressors present alternatives that are capable of effectively overcoming all the problems listed (provided these assumptions are valid) even if the researcher does not have good instrumental variables that are external to the model

Originality/value – The paper discusses and illustrates a greater number of endogeneity problems, showing how they are addressed by different estimators for panel data, using less technical and more accessible language for researchers not yet initiated in the intricacies of estimating dynamic models for panel data.

Keywords – Corporate Finance; Econometrics; Panel Data; GMM



**Revista Brasileira de Gestão
de Negócios**

DOI:10.7819/rbgn.v22i0.4059

I Introduction

A large proportion of empirical studies in corporate finance use panel data, observing N firms over T time periods (typically, with a much lower T than N). The data are derived from financial statements, market quotations, and management reports, among other sources, often with the aim of relating variables and discerning to what extent an independent variable (explanatory variable or regressor) influences the behavior of the dependent variable (response variable). For example, one of the most prolific research lines in this tradition is the search to identify the determinants of firms' capital structures, examining the reasons for which some firms are relatively highly leveraged, while others use relatively more equity capital to finance their activities (e.g., Fama & French, 2002). Other areas of investigation analyze the various factors that can influence the market value, financial performance, or operational performance of firms. These factors can include the firm's capital structure, its corporate governance structure, and the characteristics of its managers, among others (e.g., Bertrand & Schoar, 2003; Himmelberg, Hubard, & Palia, 1999).

In all the examples above, the researcher is interested in discerning causal relationships between the variables of interest using real data. Traditionally, the linear regression has been the method of choice for this purpose. Of all the assumptions needed for a regression analysis to yield appropriate inferences regarding causal relationships between variables, the most important is the assumption of exogeneity of the regressors. This is the hardest to verify and the most implausible when data collected from firms are used. In practice, this assumption rules out any correlation between the explanatory variables and the error term of the postulated empirical model. If the non-correlation assumption is invalid, one or more regressors are said to be endogenous. Endogeneity of the regressors makes the estimators inconsistent and results in inappropriate inferences. The endogeneity problem in the context of corporate finance

normally derives from the existence of omitted variables, measurement errors of the variables included in the model, and/or simultaneity between the dependent and independent variables.

The main advantage of panel data regressions, which combine cross-sectional and longitudinal dimensions, is the possibility of modeling the unobserved heterogeneity (also called firm fixed effects or specific effects, supposing that the firm is the basic unit of study), representing, for example, temporally stable characteristics related to the nature of the firm's economic activity or to the quality of its management. Depending on the research context, it is possible to reduce or eliminate the endogeneity problem derived from omitted variables by eliminating the unobserved heterogeneity of the observational units. There is, however, a price to pay: in models that isolate the unobserved heterogeneity the consistency of the estimator relies on the absence of a correlation between the explanatory variables and the error term of the model at each and every point in time. This condition is known as strict exogeneity and it is often ignored in the empirical literature on corporate finance.

The assumption of strict exogeneity is necessarily violated when the model includes lags of the dependent variable, which should be quite common, as argued in this paper, given the dynamic nature of most of the phenomena of interest in corporate finance (the resulting distortion is known as short panel bias, as it is more accentuated when T is much smaller than N , which is typical of studies in this area). Even less well-known is the violation of strict exogeneity resulting from feedback effects from the response variable to the regressors. This problem, also known as dynamic endogeneity, will be frequent in studies in the area, since shocks that affect the dependent variable (e.g., indicators of investment decisions, financing, or financial performance of firms) will probably affect any determinants of these variables (i.e., regressors) in subsequent periods.

One solution to the dynamic endogeneity problem is the use of specific lags (and/or temporal differences) of the original regressors as instrumental variables, assuming zero correlation between the instruments and the model errors (i.e., sequential exogeneity assumptions). The main objective of this study is to describe this estimation strategy in a succinct and practical way, showing, by means of the theoretical discussion illustrated by an original simulation exercise, how combining it with adequate modeling of firm and time fixed effects can address not only the dynamic endogeneity problem, but also those derived from the presence of omitted variables, measurement errors, and simultaneity between dependent and independent variables.

Standing out among similar methodological papers, with a focus on finance, are Dang, Kim, and Shin (2015), Flannery and Hankins (2013), Wintoki, Linck, and Netter (2012), and Zhou, Faff, and Alpert (2014). Each one of these adopts a specific focus, with different applications and emphasizing different aspects and challenges of estimating dynamic regression models for panel data. For example, Wintoki et al. (2012) focus on the relationship between board structure and firm performance and do not use simulation to compare the performance of different estimators in terms of bias and precision. Flannery and Hankins (2013) and Zhou et al. (2014) use different simulations to compare the performance of estimators in similar empirical contexts to those found by researchers in the area of corporate finance, but in none of them do they model the possible simultaneous determination between the response variable and the regressors. Dang et al. (2015) focus on estimating the coefficient of the lagged dependent variable and assume, in their simulations, that the other regressors do not present dynamic endogeneity problems or simultaneity. Considering the complexity of estimating empirical models with observational data in corporate finance, it is not surprising that these papers sometimes reach different conclusions and recommendations, without being

able to identify a uniformly superior estimation strategy.

This study differs from the previous literature firstly because it uses Monte Carlo simulations to discuss and illustrate a greater number of endogeneity problems (i.e., feedback effects, omitted variables, measurement errors, and simultaneity), together and separately, showing how they are addressed by different estimators for panel data. In particular, this is the only study, as far as we know, to explicitly model the so-called time fixed effect, showing that its omission can introduce a relevant omitted variable bias. Secondly, this article uses less technical and more accessible language for researchers not yet initiated in the intricacies of estimating dynamic models for panel data. On the other hand, this study is less technically complex than the aforementioned ones and does not discuss the technical difficulties of applying panel data estimators when the assumptions that ensure their correction are violated, for example due to censoring of the dependent variable or the presence of autocorrelation in the model errors. Therefore, this study may serve as a complementary reference for researchers, but does not aspire to substitute other methodological guides.

The theoretical discussion suggests that endogeneity problems must often affect empirical studies with observational data in corporate finance and the simulations show that such problems can substantially undermine inferences based on estimators that are unable to adequately address them. In particular, this study warns of the possible inconsistency, in many contexts of interest, of the traditional Ordinary Least Squares (OLS), Random Effects (RE), and Fixed Effects (FE) estimators. On the other hand, certain panel data estimators based on the Generalized Method of Moments (GMM), for example the one known as System (or Blundell-Bond) GMM, are, in carefully specified models, able to address the main endogeneity concerns and thus produce more appropriate inferences even in the absence of natural experiments or of

instrumental variables that are external to the model. However, the consistency of any estimator depends on the validity of the assumptions underlying it. Although the assumptions of the aforementioned GMM estimators are less restrictive and more plausible than those of more traditional estimators, the econometric literature shows that their violation can substantially distort the inferences (Bun & Sarafidis, 2015; Dang et al., 2015). In addition, data limitations and specification problems in the regressions may result in substantial finite sample bias (i.e., when using relatively small samples. See, for example, Windmeijer, 2018; Bun & Sarafidis, 2015).

The paper is structured as follows: section 2 discusses the main causes of the endogeneity problem in the context of corporate finance and the use of instrumental variables as a generic solution to this problem; section 3 discusses the regression methods for panel data most commonly used in empirical research in corporate finance and employed in our simulations; section 4 presents and discusses our main results; and section 5 concludes the paper.

2 Sources of Endogeneity and Instrumental Variables

Consider the following linear model:

$$y_i = \alpha + \beta x_i + \varepsilon_i, i = 1, \dots, N, \quad [1]$$

in which i corresponds to the i -th firm in a random sample containing N firms, y is the response variable, x is the regressor of interest, and ε is the error term. Suppose that the parameter β represents the causal effect (linear, in this example) of x (e.g., size of the firm, its corporate governance practices, leverage, etc.) on y (e.g., financial performance, board structure, etc.). In order to estimate β consistently (i.e., any bias converges to zero as N increases), a fundamental assumption is that of a non-correlation between x and ε , in which case x would be defined as an exogenous regressor. However, the exogeneity assumption cannot be easily verified, since, unlike x and y ,

ε is not directly observable. The fundamental causes that lead to its violation are well known and discussed below.

2.1 Sources of endogeneity

2.1.1 Omitted variables

Perhaps the most common (or most evident) cause of endogeneity in regression models is the omission of variables simultaneously correlated with the included regressors and with the response variable. In equation (1), the problem can be represented by a variable w that influences the behavior of y and of x at the same time. Its omission in (1) means that w will be incorporated into the error ε , causing some correlation between ε and the variable of interest x and introducing bias into the estimation of β . One standard solution to the problem would be to include w among the regressors, thus expanding the original model, as shown below:

$$y_i = \alpha + \beta x_i + \delta \omega_i + \varepsilon_i. \quad [2]$$

In this case, w would be considered a control variable. The inclusion of control variables (e.g., $\omega_1, \dots, \omega_k$) in regressions has been a preferential way of avoiding possible endogeneity problems in the empirical studies in corporate finance (w can also be some transformation of x , for example x^2 or x^3 , aiming to capture non-linear relationships between x and y , for example). This strategy will not work if w is intrinsically unmeasurable or if the researcher does not have enough information to measure it reliably. Unfortunately, this can be expected to occur in a good portion (if not most) of the empirical studies in this area of research.

It is not hard to think of examples of unobservable (or unmeasured) omitted variables in the context of corporate finance. For example, w could represent the ability of managers, elements of the organizational culture, or competitive advantages of the firm possibly correlated with y and x . Even potentially

measurable variables such as the firm's market power, which could simultaneously influence its financial performance, market value, financing structure, growth opportunities, and corporate governance practices, among other indicators of interest, are often ignored in empirical studies due to the unavailability of data or the difficulty of computing proxy variables that effectively capture the phenomenon.

2.1.2 *Measurement errors*

In studies with observational data on firms, it is reasonable to suppose that both y and x may be measured with some degree of imprecision, caused both by recording errors (e.g., typos or rounding) and by divergence between the construct that one wishes to observe and the proxy that is actually available. Generically, we can represent the problem via the equation:

$$x_i = x_i^* + e_i, i = 1, \dots, N, \quad [3]$$

in which x_i is the variable actually observed, x_i^* is its "true" value, and e_i is the measurement error, or noise. Similar reasoning would apply to the y variable.

Normally, the theoretical arguments that guide the formulation of the empirical models postulate certain relationships between constructs (e.g., value, performance, size, quality of corporate governance practices, etc.) that often do not correspond exactly to the indicators observed by the researcher. In other words, suppose that the model that one would like to estimate is:

$$y_i^* = \alpha + \beta x_i^* + \varepsilon_i, \quad [4]$$

but that only measures y_i and x_i , possibly measured with an error, are available. This is of course a common difficulty in many empirical studies in the field of corporate finance, and its effects over the resulting estimates depend on assumptions regarding the behavior of the measurement errors.

Suppose that only x is measured with an error and that the model that one would like to estimate is $y_i = \alpha + \beta x_i^* + \varepsilon_i$. Since x_i^* is unobservable, the equation actually estimated, substituting equation (3) in the equation above, will be:

$$\begin{aligned} y_i &= \alpha + \beta x_i^* + \varepsilon_i \\ &= \alpha + \beta(x_i - e_i) + \varepsilon_i \\ &= \alpha + \beta x_i + u_i, \end{aligned} \quad [5]$$

so that $u_i \equiv \varepsilon_i - \beta e_i$ is the error term of the model actually estimated. In this case, β will be consistently estimated if u and x are non-correlated. For this, it is necessary that no correlation exists between e and x , nor between the measurement error e and x . Unfortunately, even if the first assumption is valid, in many cases the second will not be. As an illustration, x may be the firm's observed market value, x^* the portion of x determined by the fundamentals of the business evaluated by the investors, and e the portion of the price due to various forms of noise, including speculative movements. The pricing errors aggregated in e may be independent of the firm's fundamentals, but are probably positively correlated with the market value observed by the researcher.

When e and x are correlated, the traditional estimators for the parameters of equation (5) become inconsistent. More specifically, the estimated value for the β coefficient would probably be lower in magnitude than its true value (that which would be obtained if x were measured without any error), a phenomenon known as attenuation bias. However, if several regressors contain measurement errors that are correlated with their observed values the direction of the resulting inconsistency is usually undetermined (Greene, 2000). Analogous reasoning applies to a measurement error in y correlated with x . In any case, the resulting inconsistency is similar to the one produced by omitted variables (for a detailed discussion, see Roberts & Whited, 2013).

2.1.3 Simultaneity

A common source of endogeneity problems in corporate finance research is the probable simultaneous determination of most firm-level outcomes and characteristics. In fact, considering the complex interdependency of corporate decisions, it can be argued that this should be a first order concern for empirical researchers in the area. One example is the relationship between the leverage and market value of firms. Different theoretical arguments suggest that measures of market value, such as proxies for future investment opportunities, can contemporaneously influence the financing policy of firms (Fama & French, 2002). At the same time, other arguments suggest that the degree of leverage can have an influence over the organization's performance, for example by reducing its available cash, which could otherwise be used inefficiently by self-interested managers, thus partly contributing to the determination of the firm's market value (McConnell & Servaes, 1995; Stulz, 1990). Similar reasoning can be applied to many other corporate variables, making the direction of the expected causal relationships ambiguous.

Possible simultaneity (also known as simultaneous determination or reverse causality) in the relationship between y and x , so that both variables can be considered independent or dependent in relation to each other, will introduce some correlation between the regressor and the model error, again making the estimators of β that ignore the problem biased and inconsistent.

2.2 Instrumental variables and quasi-experiments

The generic solution for any endogeneity problem, whether it is produced by measurement errors, omitted variables, or simultaneity, is the use of valid instrumental variables. Returning to the initial model $y_i = \alpha + \beta x_i + \varepsilon_i$, the x variable will be endogenous if it is correlated with ε . This problem will make it impossible to

consistently estimate the parameter of interest β , unless there is another variable z that is, at the same time, correlated with x and non-correlated with ε . Therefore, with respect to the model above, z would be an exogenous variable. In this case, one possibility is to implement an estimation in two stages as illustrated below. First, the parameters of the model that relates x and z are estimated:

$$x_i = \delta_1 + \delta_2 z_i + v_i, i = 1, \dots, N, \quad [6]$$

assuming that $\delta_2 \neq 0$. Then, the estimated parameters ($\hat{\delta}_1$ and $\hat{\delta}_2$) are used to construct a variable (\hat{x}_i) resulting from the projection of x in z so that $\hat{x}_i \equiv \hat{\delta}_1 + \hat{\delta}_2 z_i$. Therefore, \hat{x}_i corresponds to the adjusted or predicted values for this first linear regression.

In the second stage, the original variable x is substituted by \hat{x} and equation (7) below is estimated:

$$y_i = \alpha + \beta \hat{x}_i + \varepsilon_i, i = 1, \dots, N. \quad [7]$$

Since there is no correlation between z and ε , there will also be no correlation between \hat{x} and ε . In fact, \hat{x} can be understood as the portion of x that is not correlated with ε . When more than one exogenous instrument for x is available they can be included as additional regressors in equation (6). Despite the simplicity of this identification strategy, the great challenge for researchers is to find a valid instrument or set of instruments sufficiently correlated with the endogenous variables. This difficulty is worsened because, although the first assumption, of a significant correlation between the instruments and the endogenous regressor, is verifiable, the second, of a non-correlation between them and the error term of the model, is not, since the error is not directly observable. Larcker and Rusticus (2010) discuss, in the context of accounting research, which is similar to that of research in corporate finance, the main problems and challenges of the identification

strategies that use instrumental variables that are external to the model, highlighting the problem of weak instruments and the probable endogeneity of many instruments proposed in the literature.

With the aim of increasing the credibility of its identification strategies, a growing portion of the corporate finance literature uses instrumental variables derived from particular contexts, generically called natural experiments or quasi-experiments. Most of these studies explore apparently exogenous particularities or events (therefore, not influenced by the corporate variables of interest themselves), including changes in laws and regulations imposed on a set of firms. Besides constructing instrumental variables, quasi-experimental contexts are amenable to the use of other strategies for identifying causal effects, including event studies, regression discontinuity designs, difference-in-differences models, and propensity score matching (see, for example, Roberts & Whited, 2013; Angrist & Pischke, 2008). In some cases, two or more of these strategies are used simultaneously with the aim of mitigating endogeneity concerns. For example, Black and Kim (2012) investigate the influence of board structure over the market value of South Korean firms by focusing on a regulatory change that applied exclusively to large-sized firms. Based on this natural experiment, these authors employ instrumental variables estimation, regression discontinuity design, and difference-in-differences modeling, and they conduct an event study.

Studies in corporate finance that employ instrumental variables and/or quasi-experiments also often use estimation methods for panel data as part of their empirical strategy (e.g., Black & Kim, 2012). However, the main attraction of the estimation procedures discussed in the next sections is the possibility of mitigating endogeneity problems of the regressors even in

the absence of instruments that are external to the model and of quasi-experimental contexts, this absence being common in most empirical studies in corporate finance.

3 Regression Methods for Panel Data

Adding the longitudinal dimension to general equation (1), we represent the empirical model of interest as follows:

$$y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}, i = 1, \dots, N, t = 1, \dots, T. [8]$$

The only difference between (1) and (8) is that, now, N firms are observed over T time periods, so that subscript i and t represent, respectively, the i -th firm and the t -th time period. Below, we discuss the modeling possibilities offered by panels and their potential benefits in controlling endogeneity problems. In general, the procedures presented below are appropriate for short panels, understood as those in which N is much bigger than T , as is the case of most of the samples available to corporate finance researchers. Thus, all the asymptotic results applicable to the discussion below are based on the assumption that T is fixed and $N \rightarrow \infty$ (or, less formally, T is fixed and N is sufficiently large).

3.1 Unobserved heterogeneity

One of the most interesting possibilities offered by samples arranged in a panel is the explicit modeling of variables that are not observed by the researcher (whether due to lack of information, or because these variables are intrinsically unobservable). This new component can be represented as a break-down of the error term of equation (8), in the form of $\varepsilon_{it} = \eta_i + u_{it}$, resulting in the extended model below:

$$y_{it} = \alpha + \beta x_{it} + \eta_i + u_{it}, i = 1, \dots, N, t = 1, \dots, T, [9]$$

in which η_i represents the unobserved heterogeneity of the firms in the sample and u_{it} is the error term of the model. The only restriction on the behavior of η_i is that it should vary only between firms and not over time. In practice, this means that η_i captures each and every unobserved heterogeneity associated with firm i that is invariant over the course of the sampling period. In the context of corporate finance, this can include elements of the firm's organizational culture, the ability or intellectual capital of its collaborators, its innovation capacity, as well as other competitive advantages and idiosyncrasies, including ones linked to the nature of its business activity, as long as they are stable over time or, at least, over the sampling period.

Depending on the method used to estimate the parameters of model (9), the inclusion of η_i may help to mitigate or eliminate the problem of omitted variables, which is so common in many empirical contexts of interest in corporate finance, effectively complementing the traditional inclusion of control variables (in this case, only control variables that vary over the sampling period would need to be included).

The estimation of models containing η_i can be run in various ways, depending on the research objectives and the assumptions adopted by the researcher. The different procedures are often grouped into two categories: Random Effects (RE) and Fixed Effects (FE)¹. In both cases, the consistent estimation of β fundamentally depends on the assumption of a non-correlation between the error u_{it} and the regressor of interest x observed at any point in time. Therefore, not only the non-correlation between u_{it} and x_{it} , but also between u_{it} and x_{i1}, \dots, x_{iT} is assumed. The RE approach, however, uses the additional assumption of a non-correlation between x_{i1}, \dots, x_{iT} and the specific effect η_i . Regarding the identification of the β parameter, this can be considered as the fundamental difference between the two approaches². If the assumption of a non-correlation between x and h is deemed unrealistic, the FE procedures will, in principle, be more appropriate.

3.2 The strict exogeneity assumption and feedback effects

The fundamental assumption for correctly estimating the parameters of models with unobservable heterogeneity using the traditional FE and RE procedures may be more restrictive than it appears and warrants specific examination. To facilitate the explanation, statements regarding the correlation between errors and regressors will be substituted by statements regarding the conditional expectation of the errors. Thus, the fundamental assumptions for estimating the parameters of equation (8) using the FE and RE procedures can be formalized as:

$$E(u_{it}|x_{i1}, x_{i2}, \dots, x_{iT}, \eta_i) = 0, \quad [10]$$

in which $E(\cdot)$ is the expected value operator. The expression above is known as the assumption of strict exogeneity of the regressors and is a sufficient condition for the non-correlation between u_{it} and x_{i1}, \dots, x_{iT} . The strict exogeneity assumption rules out any correlation between the current errors and past, current, or future values of the explanatory variables. Although this is an acceptable assumption in some research contexts, in many others it is unrealistic.

Consider, as an illustration, a typical corporate finance model with the degree of firm leverage being explained by its profitability and by its market value. The error term of this regression will capture all the shocks that may contemporaneously affect the degree of leverage, for example a business strategy overhaul that implies, among other things, the immediate reorganization of the firm's financing structure. Even if such a change does not influence the firm's profitability and market value contemporaneously, it is quite likely that it will be correlated with their future values. This phenomenon is known as a feedback effect from the response variable to the regressors, in the sense that, returning to the example, changes in the degree of leverage may affect the organization's future profitability

and market value. If there is such feedback, the assumption of strict exogeneity will not be met, making the traditional FE and RE estimators inconsistent.

In fact, in light of the interdependency of corporate decisions, it seems reasonable to expect some degree of feedback from the dependent variable to the regressors in almost all empirical contexts of interest to corporate finance researchers. This phenomenon, which is often ignored in empirical studies that use panel data, is well discussed by Wintoki et al. (2012). In their paper, the authors refer to the problem as dynamic endogeneity and offer examples of its occurrence in the context of studies that investigate the relationship between firm performance and corporate governance.

The problem described above may be solved using any FE or RE estimators adapted to accommodate instrumental variables, provided valid strictly exogenous instruments are available. Alternatively, some procedures, presented below, enable the consistent estimation of models with unobserved heterogeneity using instruments based on lags of the original regressors and much less restrictive assumptions than the one formalized in (10).

It is important to observe that models that ignore unobserved heterogeneity, of the type $y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$, whose parameters are typically estimated by OLS applied to panel data (also known as Pooled OLS), need as a fundamental assumption the contemporaneous non-correlation between the errors and regressors. A sufficient condition is represented by equation (11):

$$E(\varepsilon_{it}|x_{it}) = 0. \quad [11]$$

This assumption is much less restrictive than that of strict exogeneity. That is, in this context the presence of feedback effects will not make the regressor endogenous. On the other hand, it is clear that the assumption in (11) will be violated if there is an unobserved effect η_i correlated with

the regressors contained in ε_{it} .

3.3 Procedures based on the generalized method of moments

The discussion above suggests that explicitly modeling the unobserved heterogeneity of firms is desirable in many corporate finance research settings. However, the most commonly employed methods for estimating models of this type, often classified as RE or FE estimators, require the regressors to be strictly exogenous, an assumption that is probably very restrictive in studies that use firm data and that will be violated if there is feedback from the response variable to the regressors. Naturally, the other potential sources of endogeneity problems, presented in the previous sections, can also contribute to violating this assumption.

A natural solution to this problem is to use instrumental variables that are external to the model of interest. It is theoretically possible, for example, to find strictly exogenous instruments for each one of the regressors suspected of endogeneity. In practice, however, variables with these characteristics and that also present a strong correlation with the regressors are rarely available in corporate finance studies. The methods described in this section, on the other hand, enable the use of instruments that are only sequentially exogenous, based, for example (but not necessarily), on adequate lags of the original regressors themselves.

Consider again the model shown in (9). Suppose that x is correlated (through a feedback effect) with the past values of the error term ($u_{it-1}, u_{it-2}, \dots, u_{i1}$), but that it is not correlated with its current or future values. A sufficient condition for this last assumption can be expressed in the form:

$$E(u_{it}|x_{i1}, x_{i2}, \dots, x_{it}, \eta_i) = 0. \quad [12]$$

In this case, x is assumed to be sequentially exogenous, as opposed to the more restrictive assumption of strict exogeneity formalized by

equation (10) (Wooldridge, 2010). The idea of sequential exogeneity can be naturally extended to accommodate any lags or leads of the regressors that are supposedly non-correlated with the errors. The simultaneous determination of the regressors and of the response variable, for example, can result in some correlation between x_{it} and u_{it} . In this case, assumption (12) will not be valid, but the assumption

$$E(u_{it}|x_{i1}, x_{i2}, \dots, x_{it-1}, \eta_i) = 0, \quad [13]$$

will be appropriate if there is no correlation between the regressors and the future values of the error term of the model. Similar endogeneity problems can result from the presence of measurement errors in x_{it} and their solution may also involve assumptions of sequential exogeneity of the regressors³.

Various estimation methods that are appropriate for short panels and that use sequentially exogenous variables as instruments are available and are sometimes classified into two groups: estimators of Instrumental Variables and estimators based on the Generalized Method of Moments (GMM). These methods have been developed with a focus on estimating dynamic

models, meaning empirical models that include among the regressors one or more lags of the response variable, typically only the first lag. In other words, in a formulation such as the one shown in (9), y_{it-1} would be included among the regressors and, by definition, y_{it-1} is not a strictly exogenous variable. However, the methods discussed here are equally valid for static models such as the one shown in (9), that is, formulations that do not include lags of y_{it} among the regressors. A good introduction to this literature is offered by Bond (2002). Among the various methods developed for panels that are able to incorporate instrumental variables, two stand out due to their efficiency and flexibility for accommodating different patterns of behavior of the variables of interest. The first is a procedure developed by Arellano and Bond (1991) and known as the Arellano-Bond estimator or First-differencing GMM (GMM-Dif).

This procedure first transforms the variables of the model with the aim of eliminating the unobserved heterogeneity. The transformation normally applied consists of computing the difference of each variable with relation to its first lag. Applying this transformation to model (9), equation (14) is obtained:

$$\Delta y_{it} = \beta \Delta x_{it} + \Delta u_{it}, i = 1, \dots, N, t = 1, \dots, T, \quad [14]$$

with $\Delta y_{it} \equiv y_{it} - y_{it-1}$, $\Delta x_{it} \equiv x_{it} - x_{it-1}$ and $\Delta u_{it} \equiv u_{it} - u_{it-1}$. This procedure eliminates the unobserved heterogeneity, since $\Delta \eta_i = 0$. This transformation, known as first differencing, is classified as a FE-type procedure and makes no assumption regarding the correlation between η_i and x_{it} . Other transformations capable of eliminating the unobservable component η_i are also possible in this context, for example transformation via orthogonal deviations, as described by Arellano (2003).

After eliminating the unobserved heterogeneity, the procedure estimates the

parameters in (14) by GMM, exploiting the exogeneity assumptions assumed by the researcher. For example, if there is reason to believe that there are significant feedback effects from y to x , it cannot be assumed that x is strictly exogenous because there will be some correlation between u_{it} and $x_{it+1}, x_{it+2}, \dots, x_{iT}$ (that is, the errors influence the future values of x). However, if it is reasonable to assume that there are no simultaneity problems, omitted variables (besides those captured by η_i), or measurement errors that cause a correlation between u_{it} and current and past values of x , it can be assumed that this regressor is

sequentially exogenous. More specifically, in this case, it is said that x is a 'predetermined' variable (Arellano, 2003). Under this assumption, the estimator can use the following orthogonality (or non-correlation) conditions, generically called moment conditions:

$$E(x_{it-s}\Delta u_{it}) = 0, s \geq 1 \quad [15]$$

Using the transformed errors Δu_{it} , the expression above simply reflects the assumption of a non-correlation between u_{it} and $x_{it}, x_{it-1}, \dots, x_{i1}$ (under the trivial assumption that $E(u_{it}) = 0$).

If, however, in addition to the feedback effects, there is, for example, simultaneity in the relationship between y and x , there will be a contemporaneous correlation between u and x and assumption (15) will be inadequate. In this case, in econometric jargon, x will be an 'endogenous' variable and no longer predetermined. In fact, however, x will not be completely endogenous to the extent that its lags are not correlated with the model error. In other words, despite the jargon, x may still be sequentially exogenous and, in this case, the GMM-Dif estimator can use the following moment conditions:

$$E(x_{it-s}\Delta u_{it}) = 0, s \geq 2 \quad [16]$$

In practice, these orthogonality conditions mean that the estimator will use all of the suitable lags of x as instrumental variables, that is, variables assumed to be uncorrelated with the error term of the model. Based on this strategy and following a similar procedure to the one described in section 2.2 (although more complex than it), the coefficient of interest β is estimated.

Many moment conditions that are different from those represented by (15) and (16) can be naturally accommodated by the GMM-Dif estimator, which enables the use of not only any past or future values of x as instruments but also variables that are external to the model and

fulfill the assumptions described in section 2.2. Naturally, in the particular case in which the only relevant source of endogeneity is the presence of unobserved heterogeneity, the GMM-Dif will use x as an instrument for itself.

Blundell and Bond (1998) present the final version of an important extension of the First-differencing GMM, known as System GMM (GMM-Sys). This method uses the same moment conditions described above and adds others, thus increasing the efficiency and performance in finite samples of the estimator (Blundell, Bond, & Windmeijer, 2000). Continuing the previous example, if condition (16) is valid, the following additional moment conditions can be exploited by the system estimator:

$$E[\Delta x_{it-1}(\eta_i + u_{it})] = 0 \quad [17]$$

Unlike what is observed in (16), the first difference transformation here is applied to the regressors, which multiply the non-transformed error. This method imposes the additional assumption of a non-correlation between Δx_{it-1} (or, more generically, Δx_{it}) and η_i . This last assumption is not as restrictive as it appears because it allows for an arbitrary correlation between the regressors and unobserved heterogeneity. It only requires that this correlation does not change between one particular point in time and the next, which is often acceptable, given the nature of the specific effect η_i :

$$E(\Delta x_{it}\eta_i) = 0 \Rightarrow E(x_{it}\eta_i) = E(x_{it-1}\eta_i). \quad [18]$$

Blundell and Bond (1998) show that the non-correlation between Δx_{it} and η_i will be ensured if the stochastic process that generates x_{it} is stationary. This is a sufficient condition and one that can be tested, but it is not necessary. Weaker sufficient conditions, relating to the behavior of the initial values of the time series (x_{i1} , in the example) are discussed by Blundell and Bond (1998, 2000) and Bond (2002).

In short, the more advanced panel estimation procedures based on GMM enable the researcher to resort to less restrictive assumptions than those that are necessary to ensure the consistency of the estimators traditionally used in empirical corporate finance research. In addition, they are particularly useful when the researcher

does not have instrumental variables that are external to the model and/or quasi-experimental contexts.

3.4 Time fixed effects

A second extension of the basic panel data model that relates x and y is:

$$y_{it} = \alpha + \beta x_{it} + \eta_i + \lambda_t + v_{it}, i = 1, \dots, N, t = 1, \dots, T. \quad [19]$$

Now, the original error term is broken down into three components: $\varepsilon_{it} = \eta_i + \lambda_t + v_{it}$, where η_i is the unobserved heterogeneity and v_{it} is the idiosyncratic error term. The novelty in (19) is λ_t , which represents the so-called time fixed effects. This component only varies in time and not between firms, capturing each and every shock in y that has simultaneously affected all the firms in the sample.

It is easy to show that the explicit modeling of λ_t , which is often ignored, can be quite important in empirical studies in corporate finance. Practically any response variable of interest in this area can be significantly affected by macroeconomic shocks, for example unexpected variations in inflation, interest, or exchange rates or significant variations in the country's fiscal policy. For example, the performance of all (or almost all) non-financial firms will be negatively affected if there is a sudden rise in the basic interest rate, making credit more expensive and reducing demand. Thus, if y represents financial performance, the common component of the negative shock caused by the rise in interest rates will be captured by λ_t . In fact, λ_t captures the impact on y (common to all firms in the sample) of a potentially wide set of macroeconomic shocks

occurring in period t (over the course of a year, for example). Even if the same macroeconomic shocks do not have any influence over x , ignoring the λ_t component (and therefore leaving it within the error term of the model) may adversely affect the estimation of the coefficient standard errors (Fama & French, 2002). The problem will be greater, however, if λ_t is correlated with x . In this case, λ_t will be an omitted variable, rendering the typical estimators for panel data (including all those previously mentioned) inconsistent. This will likely happen if x represents, for example, the firm's size (measured by its net sales), its degree of leverage, profitability, or managers' equity holdings.

Fortunately, it is perfectly feasible to isolate the potentially relevant impact of λ_t and the most practical way of doing this is to include in the regression a set of time indicator variables (d_t , $t = 1, \dots, T$), so that $d_t = 1$ in period t and $d_t = 0$ otherwise (naturally, this variable rules out subscript i because it does not vary between firms). Therefore, the model actually estimated (by any of the methods discussed previously) will be (excluding d_1 from the equation to avoid perfect collinearity of the regressors, since the model includes an intercept):

$$y_{it} = \alpha + \beta x_{it} + \eta_i + \sum_{t=2}^T d_t + v_{it}, i = 1, \dots, N, t = 1, \dots, T. \quad [20]$$

3.5 Dynamic models

The models outlined up to here ignore the possible direct influence of past values of the response variable over its current values. However, many of the indicators of interest in corporate finance present strongly inertial behavior (e.g., governance practices, financial performance, leverage, turnover), suggesting that the specification of static models may not be adequate.

Different arguments can explain such behavior. For example, Wintoki et al. (2012) suggest that the high persistence of firms' profitability, documented in various empirical studies (Glen, Lee, & Singh, 2001; Waring,

1996) reflects, to some extent, unobserved variables such as managerial ability (which may have some variation in time, and for this reason is not perfectly captured by the fixed effect). In addition, it is common to observe some behavior of regression to the mean in corporate variables, causing a negative correlation between the current values of these variables and their subsequent variations. In fact, this partial adjustment movement towards equilibrium values is expected, for example, by different theories of capital structure that suggest the existence of an optimal financing structure for each firm (Fama & French, 2002; Frank & Goyal, 2003).

To explicitly model this dynamic component we can extend (19):

$$y_{it} = \alpha_0 + \alpha_1 y_{it-1} + \beta x_{it} + \eta_i + \lambda_t + v_{it}, i = 1, \dots, N, t = 1, \dots, T. \quad [21]$$

If the correct model is represented by (21), with $\alpha_1 \neq 0$, the omission of y_{it-1} in the regression will make the estimator of β inconsistent if y_{it-1} (which will be included in the error term of the estimated model) is correlated with x_{it} . A sufficient condition for this to occur is that x is time-persistent, so that there is a significant correlation between x_{it} and x_{it-1} . Naturally, an even more direct source of inconsistency, in this case, would be the existence of feedback from y to x , as discussed in section 3.2.

One indication of the inadequacy of the static specification is the presence of a significant autocorrelation in v_{it} , which can be empirically verified by the researcher using autocorrelation tests of the residuals of the original static regression. In many cases, the inclusion of the first lag of the response variable among the regressors is enough to capture this phenomenon, but, in theory, other lags may also be relevant to account for the dynamic behavior of y (e.g., $y_{it-1}, y_{it-2}, \dots, y_{it-p}$).

Model (21) will not be adequately estimated by any procedure that needs the assumption of strict exogeneity of the regressors,

as is the case of the traditional FE and RE estimators, since, by definition, y_{it-1} is not a strictly exogenous variable. Such an assumption, in this model, would imply a non-correlation between y_{it} and y observed at any point in time. Therefore, it would also require a non-correlation between y_{it} and y_{it} , which is impossible by construction. However, if the regressors are sequentially exogenous, the parameters of (21) can be consistently estimated by the GMM-based methods presented in section 3.3.

4 Results of the Regressions with Simulated Data and Performance of the Estimators

This section presents the procedures for building simulated panel samples with similar characteristics to those available to corporate finance researchers. Next, we present some results of regressions employing simpler and more advanced estimators applied to the simulated samples, enabling a comparison of their relative performance and an evaluation of the fit of the different estimation strategies to the data

generated. As this laboratorial analysis synthesizes key and important aspects of the data typically used in empirical studies in corporate finance, it can offer a methodological guide for researchers in the area, on one hand highlighting some of the biggest concerns they should pay attention to and, on the other hand, offering possible solutions.

4.1 General model of the simulation

The aim of this simulation analysis is to evaluate the performance of different estimation

strategies applied to samples of artificial data with similar characteristics to those actually used by corporate finance researchers in their empirical studies. For this, we use Monte Carlo procedures to generate sets of random samples based on models that synthesize the aforementioned characteristics in the most complete way possible.

The general model of the simulation is quite similar to the one shown in (21) (excluding, for simplicity and without loss of generality, the intercept α_0), where v_{it} is its random error term:

$$y_{it} = \alpha y_{it-1} + \beta x_{it} + \eta_i + \lambda_t + v_{it}, i = 1, \dots, N, t = 1, \dots, T. \quad [22]$$

It captures various potentially relevant characteristics of processes of interest to corporate finance researchers, including the dynamic behavior of the response variable (represented by y_{it-1}), the unobserved heterogeneity of the firms (η_i), and the influence of unobserved macroeconomic factors⁴ (λ_t).

Just as important as modeling the behavior of the response variable, however, is modeling the behavior of the regressor of interest x , so that the analysis contemplates several potential endogeneity problems capable of impeding the consistent estimation of the parameters of interest α and β . The general model for x is shown below (where e_{it} is its random error term):

$$x_{it} = \rho x_{it-1} + \tau \eta_i + \phi \lambda_t + \theta_1 v_{it} + \theta_2 v_{it-1} + e_{it}, i = 1, \dots, N, t = 1, \dots, T. \quad [23]$$

Model (23) allows x to exhibit some degree of temporal persistence (as observed in practice in many corporate variables) and contemplates all the endogeneity problems discussed in the previous sections, as we explain below.

The problem of omitted variables related to unobservable time-invariant firm characteristics is represented by $\tau \eta_i$ and will exist if $\tau \neq 0$. It will be more pronounced the larger this parameter is. Similarly, the problem of omitted variables related to unobservable time effects (e.g., macroeconomic shocks) is represented by $\phi \lambda_t$ and will be proportional to the value of the ϕ parameter. In turn, the endogeneity of x caused by feedback effects from y to x (also called dynamic endogeneity, as discussed in section 3.2) is captured by $\theta_2 v_{it-1}$ (y_{it-1} could be used instead

of v_{it-1} with similar results) and its magnitude will depend on the value associated with θ_2 . The possible (and probable, in many empirical contexts of corporate finance) simultaneous determination of y and x is captured by $\theta_1 v_{it}$, since the phenomenon of reverse causality will produce some contemporaneous correlation between v and x . Finally, both $\theta_1 v_{it}$ and $\theta_2 v_{it-1}$ can also account for the possible endogeneity caused by measurement errors in x or y or omitted variables that vary over time and between firms.

The construction of simulated samples based on models (22) and (23) enables us to analyze with precision the combined effects of different endogeneity problems applicable to empirical studies in the field of corporate finance.

Specifically, this computational exercise allows us to highlight the most critical challenges to consistently estimating the parameters of interest in regressions with observational data, as well as to suggest strategies for addressing them. To achieve these objectives, however, it is important to break down the general model into relevant particular

cases, thus isolating specific problems, as will be discussed in the following sections.

The complete model initially used to generate the simulated panels of this study is presented below in more detail, including the parameters chosen by the researchers for the purpose of illustration:

$$\begin{aligned} y_{it} &= \alpha y_{it-1} + \beta x_{it} + \eta_i + \lambda_t + v_{it} \\ x_{it} &= \rho x_{it-1} + \tau \eta_i + \phi \lambda_t + \theta_1 v_{it} + \theta_2 v_{it-1} + e_{it}, \end{aligned} \quad [24]$$

in which $\eta_i \sim N(0, \sigma_\eta^2)$, $\lambda_t \sim N(0, \sigma_\lambda^2)$, $v_{it} \sim N(0, \sigma_v^2)$, $e_{it} \sim N(0, \sigma_e^2)$. In this study, we assume: $\sigma_\eta^2 = \sigma_\lambda^2 = \sigma_v^2 = \sigma_e^2 = 1$, $\beta = 1$, $\tau = 0.7$, $\alpha = 0.5$, $\theta_1 = 0.6$, $\theta_2 = 0.5$, $\rho = 0.5$ and $\phi = 0.5$.

As shown in (24), we assume that η_i , λ_t , v_{it} , and e_{it} are random variables that follow a standard normal distribution. However, this choice does not imply a loss of generality because the estimation procedures employed below are asymptotically robust to deviations from normality. The original programming code was developed for Matlab and used to generate the samples according to the system of equations in (24).

4.1.1 Performance of the estimators based on the general model of the simulation

After generating the data based on model (24), with $N = 500$, $T = 8$ and 1000 replications, we estimated, for each one of the 1000 samples, the parameters of interest α and β based on five different estimation methods. Specifically, we used the traditional OLS estimator, the RE and FE estimators, as well as the GMM-based methods (GMM-Dif and GMM-Sys). All the estimation procedures were implemented in the Stata statistical package, using the 'xtabond2' function. All Matlab as Stata codes are available from the authors upon request.

The results of the estimation of the general model are reported in Table 1. Although the model described in (24) is plagued by different

endogeneity problems, the estimation by OLS is only capable of avoiding the omitted variables bias caused by λ_t by including among the regressors a set of time dummy variables (see section 3.4). The other problems are forcibly ignored, resulting in substantial bias in the estimation of β , considering that the true value of the parameter is 1 and the mean value of the 1000 computed estimates is equal to 1.3678 (with a minimum of 1.3292 and maximum of 1.4113). The distance between the true value and the one obtained by the estimator is also reflected in the high root mean squared error (RMSE) associated with the estimator of the β parameter. The RMSE of the estimators of β is computed by the following equation:

$$RMSE = \frac{\sum_{j=1}^S (\hat{\beta}_j - \beta)^2}{S}, \quad [25]$$

in which $\hat{\beta}_j$ is the estimate of this parameter in the j -th simulated sample (of a total of S samples). In Table 1, $S = 1000$ and $\beta = 1$.

A bias of this magnitude would be economically relevant if the data corresponded to financial information from real firms. For example, based on the average of the estimates, shown in Table 1, a researcher employing the OLS

estimator could infer that, all else equal, a change of one unit in x would result in an expected change of approximately 1.4 unit in y , a distortion of 40% in relation to the correct inference. Naturally, the values presented here are mere illustrations. In real life, the magnitude of the problem would depend on several factors, including sampling variation and the magnitude of the correlation between the regressors and the error term. The model estimated by OLS in Table 1 also shows a bias in the estimation of α , although it is less pronounced.

The following model, estimated by RE, produces results with a similar bias to that reported for the OLS estimator. Although this procedure explicitly includes the unobserved heterogeneity (η_i), it assumes that it is not correlated with the

regressors. In addition, this procedure is incapable of dealing with other sources of endogeneity, such as feedback effects, measurement errors of the regressors, and simultaneity.

The result of the following model also shows a substantial bias in the FE estimator. Although this procedure is more robust than the previous ones, enabling an arbitrary correlation between η_i and the regressors, its validity fundamentally depends on the assumption that the regressors are strictly exogenous. The violation of this assumption in model (24), combined with the other endogeneity problems, results in a substantial bias in $\hat{\alpha}$ (greater than with the previous methods), whose true value is 0.5, as well as in $\hat{\beta}$.

Table 1
Summary of the estimations for the general model

Model	Param.	Mean	SD	Min.	Max.	RMSE
OLS	α	0,5344	0,0055	0,5152	0,5477	0,0349
	β	1,3678	0,0126	1,3292	1,4113	0,3680
RE	α	0,5167	0,0064	0,4964	0,5352	0,0179
	β	1,3802	0,0124	1,3429	1,4208	0,3804
EF	α	0,3134	0,0077	0,2835	0,3398	0,1868
	β	1,3510	0,0119	1,3158	1,3845	0,3512
GMM-Dif	α	0,4755	0,0239	0,3941	0,5481	0,0342
	β	1,0440	0,0426	0,9177	1,1645	0,0612
GMM-Sys	α	0,4908	0,0228	0,4269	0,5654	0,0245
	β	1,0356	0,0438	0,8640	1,1561	0,0564

Note. Regressions based on the system of equations (24), with population parameters: $\beta = 1, \tau = 0,7, \alpha = 0,5, \theta_1 = 0,6, \theta_2 = 0,5, \rho = 0,5$, and $\phi = 0,5$. . Number of simulated samples: 1,000. Param.: Estimated parameters; SD: Standard deviation; RMSE: Root Mean Squared Error. All regressions include time dummies and the lagged dependent variable. Estimators: Ordinary Least Squares (OLS); Random Effects (RE); Fixed Effects (FE); First-differencing Generalized Method of Moments (GMM-Dif); System Generalized Method of Moments (GMM-Sys).

The GMM-Dif estimator is able to appropriately address all the sources of endogeneity included in (24) by removing the unobserved heterogeneity and using lags of y and of x non-correlated with the error v_{it} as instrumental variables. The estimates of β are much closer to their true value when compared with the previous

methods. However, the result for $\hat{\alpha}$ is less satisfactory, and these estimates, although close to the true value of 0.5 on average, vary between 0.3941 and 0.5481.

The following model shows that the GMM-Sys estimator produces the most satisfactory results out of all those employed, with almost null



bias both for $\hat{\alpha}$ and for $\hat{\beta}$. Its advantage over GMM-Dif derives from the fact that it employs additional instruments based on assumptions of sequential endogeneity of the regressors.

4.2 Specific case 1: correlation between the regressor and the unobserved heterogeneity

Besides analyzing the general model, we investigate the behavior of various specific cases, that is, reductions of the general model that enable us to study the performance of the estimators in more particular contexts, isolating the possible endogeneity problems found in real data.

Specific case 1 only focuses on the problem of unobserved heterogeneity correlated with x , eliminating the other potential sources of endogeneity. In this case, we generate a whole new

simulation analysis after changing the parameters of the general model, that is, setting some of the population parameters to zero. For example, as specific case 1 is based on a static model, we set $a = 0$. Therefore, the general model is now reduced to:

$$\begin{aligned} y_{it} &= \beta x_{it} + \eta_i + v_{it} \\ x_{it} &= \rho x_{it-1} + \tau \eta_i + e_{it} \end{aligned} \quad [26]$$

The difference between the models presented in equations (26) and (24) is that the only source of endogeneity in the former is the correlation between η_i and x_{it} . Therefore, (26) is a much simpler model. Otherwise, its specification is identical to that of general model (24). Similar reasoning applies to the other specific cases. Table 2 reports the results of this analysis.

Table 2

Summary of the estimations for particular case 1

Model	Param.	Mean	SD	Min.	Max.	RMSE
OLS	β	1,4248	0,0147	1,3775	1,4672	0,4251
RE	β	1,3059	0,0182	1,2493	1,3565	0,3064
FE	β	1,0007	0,0164	0,9508	1,0563	0,0163
GMM-Dif	β	1,0006	0,0222	0,9159	1,0659	0,0222

Note. Regressions based on the system of equations (24), with population parameters: $\beta = 1, \tau = 0,7$, and $\rho = 0,5$. Number of simulated samples: 1,000. Param.: Estimated parameters; SD: Standard deviation; RMSE: Root Mean Squared Error. Simple regressions (x is the only regressor). Estimators: Ordinary Least Squares (OLS); Random Effects (RE); Fixed Effects (FE); First-differencing Generalized Method of Moments (GMM-Dif).

There is substantial bias in the estimators in the models estimated by OLS and RE, especially in the former, due to its incapacity to mitigate the existing endogeneity. On the other hand, the estimators based on fixed effects (FE and GMM-Dif) present quite satisfactory results, as expected. In fact, the FE estimator, in this case, presented the best performance of all, which was marginally superior to GMM-Dif. The table omits the GMM-Sys estimates because, in this simple static model, the two GMM estimators yield essentially identical results.

4.3 Specific case 2: temporal persistence of the response variable

Specific case 2 highlights the importance of including dynamic terms in the model when the response variable is highly persistent. Many empirical studies in corporate finance estimate only static models ($a = 0$). Table 3 illustrates the consequences of this potentially inadequate specification of the empirical model.

$$\begin{aligned} y_{it} &= \alpha y_{it-1} + \beta x_{it} + v_{it} \\ x_{it} &= \rho x_{it-1} + e_{it} \end{aligned} \quad [27]$$

Table 3 shows a substantial bias in all the estimators when we do not include y_{it-1} among the regressors of the model using the system of equations (27), especially the OLS estimator. It is

also interesting to note that GMM-Dif tends to underestimate the β parameter, while the others tend to overestimate it. We omit the estimates produced by RE and GMM-Sys to save space.

Table 3
Summary of the estimations for particular case 2

Model	Param.	Mean	SD	Min.	Max.	RMSE
Sta-OLS	β	1,3330	0,0247	1,2663	1,4063	0,3339
Dyn-OLS	α	0,4998	0,0088	0,4725	0,5263	0,0088
	β	1,0003	0,0158	0,9493	1,0463	0,0158
Sta-FE	β	1,1042	0,0230	1,0390	1,1684	0,1067
Dyn-FE	α	0,4064	0,0118	0,3697	0,4434	0,0943
	β	1,0144	0,0183	0,9594	1,0704	0,0232
Sta-GMM-Dif	β	0,8329	0,0182	0,7659	0,8911	0,1681
Dyn-GMM-Dif	α	0,4981	0,0178	0,4483	0,5643	0,0179
	β	0,9957	0,0229	0,9187	1,0685	0,0233

Note. Regressions based on the system of equations (27), with population parameters: $\beta = 1, \alpha = 0,5$ and $\rho = 0,5$. Number of simulated samples: 1,000. Param.: Estimated parameters; SD: Standard deviation; RMSE: Root Mean Squared Error. In the dynamic specifications (Dyn) we include the lagged dependent variable, which is absent from the static specifications (Sta). Estimators: Ordinary Least Squares (OLS); Fixed Effects (FE); First-differencing Generalized Method of Moments (GMM-Dif).

4.4 Specific case 3: feedback effects

Specific case 3 illustrates how feedback effects from y to x , captured by the term $\theta_2 v_{it-1}$, might affect estimates. Table 4 presents the results of the estimations.

$$\begin{aligned} y_{it} &= \alpha y_{it-1} + \beta x_{it} + v_{it} \\ x_{it} &= \rho x_{it-1} + \theta_2 v_{it-1} + e_{it}. \end{aligned} \tag{28}$$

As expected, the OLS estimator estimates the parameters adequately, since it does not depend on the strict exogeneity assumption and,

therefore, it is not affected by the phenomenon of dynamic endogeneity. The same does not occur, however, with the FE estimator, which depends on the strict exogeneity assumption. Since in (28) this assumption is violated, the coefficients are estimated inconsistently and the analysis shows that $\hat{\alpha}$ is the variable most affected by the problem. On the other hand, the GMM-Dif and GMM-Sys estimators, adopting the assumption that x is a predetermined variable, again present satisfactory results, with the latter estimator having a marginal advantage. We omit the estimates produced by RE to save space.

Table 4
Summary of the estimations for particular case 3

Model	Param.	Mean	SD	Min.	Max.	RMSE
OLS	α	0,4997	0,0083	0,4751	0,5240	0,0083
	β	1,0001	0,0157	0,9500	1,0470	0,0157
FE	α	0,3820	0,0108	0,3479	0,4190	0,1185
	β	0,9710	0,0173	0,9240	1,0286	0,0337
GMM-Dif	α	0,4955	0,0148	0,4428	0,5411	0,0154
	β	0,9966	0,0235	0,9182	1,0740	0,0238
GMM-Sys	α	0,4992	0,0124	0,4614	0,5317	0,0124
	β	1,0001	0,0180	0,9418	1,0570	0,0180

Note. Regressions based on the system of equations (28), with population parameters: $\beta = 1, \alpha = 0,5, \theta_2 = 0,5$, and $\rho = 0,5$. Number of simulated samples: 1,000. Param.: Estimated parameters; SD: Standard deviation; RMSE: Root Mean Squared Error. All regressions include the lagged dependent variable. Estimators: Ordinary Least Squares (OLS); Fixed Effects (FE); First-differencing Generalized Method of Moments (GMM-Dif); System Generalized Method of Moments (GMM-Sys).

4.5 Specific case 4: unobserved heterogeneity and feedback

Specific case 4 differs from case 3 in two aspects: it includes the unobserved heterogeneity in the model, maintaining the feedback effects from y to x , and removes the dynamic term. Therefore, now x will be a predetermined variable and correlated with fixed effect η_i . Consequently, there will be two simultaneous sources of endogeneity. The results of the simulations are presented in Table 5.

$$\begin{aligned} y_{it} &= \beta x_{it} + \eta_i + v_{it} \\ x_{it} &= \rho x_{it-1} + \tau \eta_i + \theta_2 v_{it-1} + e_{it} \end{aligned} \quad [29]$$

Table 5
Summary of the estimations for specific case 4

Model	Param.	Mean	SD	Min.	Max.	RMSE
OLS	β	1,3851	0,0144	1,3407	1,4439	0,3854
RE	β	1,2889	0,0186	1,2376	1,3457	0,2894
FE	β	0,9178	0,0146	0,8598	0,9678	0,0835
GMM-Dif	β	0,9971	0,0264	0,9080	1,0948	0,0265
GMM-Sys	β	1,0040	0,0226	0,9273	1,0653	0,0229

Note. Regressions based on the system of equations (29), with population parameters: $\beta = 1, \tau = 0,7, \theta_2 = 0,5$, and $\rho = 0,5$. Number of simulated samples: 1,000. Param.: Estimated parameters; SD: Standard deviation; RMSE: Root Mean Squared Error. Simple regressions (x is the only regressor). Estimators: Ordinary Least Squares (OLS); Random Effects (RE); Fixed Effects (FE); First-differencing Generalized Method of Moments (GMM-Dif); System Generalized Method of Moments (GMM-Sys).

4.6 Specific case 5: unobserved heterogeneity, measurement errors, and/or simultaneous determination

Specific case 5 is similar to case 4, but now we model x as a variable that is contemporaneously correlated with v , due,

for example, to measurement errors and/or its simultaneous determination with the response variable, a common problem in studies using firm-level data. The results are presented in Table 6.

$$\begin{aligned} y_{it} &= \beta x_{it} + \eta_i + v_{it} \\ x_{it} &= \rho x_{it-1} + \tau \eta_i + \theta_1 v_{it} + e_{it}. \end{aligned} \tag{30}$$

Table 6
Summary of the estimations for specific case 5

Model	Param.	Mean	SD	Min.	Max.	RMSE
OLS	β	1,5301	0,0109	1,4945	1,5659	0,5303
RE	β	1,4929	0,0111	1,4523	1,5230	0,4930
FE	β	1,3768	0,0126	1,3370	1,4187	0,3770
GMM-Dif	β	1,0285	0,0561	0,8353	1,1919	0,0629
GMM-Sys	β	1,0381	0,0476	0,8715	1,1782	0,0609

Note. Regressions based on the system of equations (30), with population parameters: $\beta = 1, \tau = 0,7, \theta_1 = 0,6$ and $\rho = 0,5$. Number of simulated samples: 1,000. Param.: Estimated parameters; SD: Standard deviation; RMSE: Root Mean Squared Error. Simple regressions (x is the only regressor). Estimators: Ordinary Least Squares (OLS); Random Effects (RE); Fixed Effects (FE); First-differencing Generalized Method of Moments (GMM-Dif); System Generalized Method of Moments (GMM-Sys).

Now the bias is much greater than in the previous case for the OLS, RE, and FE estimators, illustrating the substantial impact of endogeneity problems caused, for example, by questions of reverse causality between the regressors and the response variable. Since, besides this source of endogeneity, there is a correlation between x and η_i , it is not surprising that the greatest bias come from the OLS and the RE estimators. Once again, the GMM-Dif and GMM-Sys estimators produce good results, with the system estimator (GMM-Sys) having an insignificant bias and some advantage in precision.

4.7 Specific case 6: measurement errors and/or simultaneous determination without unobserved heterogeneity

Specific case 6 differs from the previous one as it focuses only on the current correlation between x and v , removing the unobserved heterogeneity from equations system (31). The results are presented in Table 7.

$$\begin{aligned} y_{it} &= \beta x_{it} + v_{it} \\ x_{it} &= \rho x_{it-1} + \theta_1 v_{it} + e_{it} \end{aligned} \tag{31}$$

Table 7

Summary of the estimations for particular case 6

Model	Param.	Mean	SD	Min.	Max.	RMSE
OLS	β	1,3314	0,0110	1,2918	1,3616	0,3316
RE	β	1,3316	0,0110	1,2925	1,3616	0,3318
FE	β	1,3765	0,0133	1,3396	1,4266	0,3767
GMM-Dif	β	1,0164	0,0435	0,8759	1,1350	0,0465
GMM-Sys	β	1,0180	0,0360	0,8701	1,1540	0,0403

Note. Regressions based on the system of equations (31), with population parameters: $\beta = 1$, $\theta_1 = 0,6$, and $\phi = 0,5$. Number of simulated samples: 1,000. Param.: Estimated parameters; SD: Standard deviation; RMSE: Root Mean Squared Error. Simple regressions (x is the only regressor). Estimators: Ordinary Least Squares (OLS); Random Effects (RE); Fixed Effects (FE); First-differencing Generalized Method of Moments (GMM-Dif); System Generalized Method of Moments (GMM-Sys).

Since the contemporaneous correlation between x and z is the only source of endogeneity in this case, the bias in the OLS, RE, and FE estimators is smaller than in the previous case, but it remains substantial. As was also expected, due to the absence of the unobserved heterogeneity, the degree of bias in the three estimators is similar. Again, the GMM-Dif and GMM-Sys estimators yield good results, with the system estimator having a slight advantage in precision.

4.8 Specific case 7: time fixed effects

In specific case 7, we highlight the importance of the time fixed effects, captured by λ_t , and their possible correlation with the regressors of the model, represented by $\phi\lambda_t$. In this case, the models are estimated with and without time dummy variables to illustrate the potential impact of the omission of these regressors in (32). The results are presented in Table 8.

$$\begin{aligned} y_{it} &= \beta x_{it} + \lambda_t + v_{it} \\ x_{it} &= \rho x_{it-1} + \phi\lambda_t + e_{it} \end{aligned} \quad [32]$$

The results of the simulation analysis show that the omission of the time dummies when there are relevant macroeconomic effects that affect x and y simultaneously can lead to substantial biasing of any of the estimators employed. In particular, we note that the GMM-Dif and GMM-Sys estimators are, respectively, the ones that present the greatest degree of bias when the problem is ignored. This result warns of the importance of including time dummies in panel data regressions in studies involving corporate variables, which are likely to be influenced by macroeconomic shocks or similar cyclical phenomena.

Table 8
Summary of the estimations for specific case 7

Model	Param.	D(time)	Mean	SD	Min.	Max.	RMSE
OLS	β	não	1,2418	0,1107	0,9737	1,5901	0,2659
	β	sim	0,9999	0,0133	0,9586	1,0698	0,0134
FE	β	não	1,3268	0,1411	0,9746	1,7401	0,3560
	β	sim	0,9998	0,0157	0,9554	1,0677	0,0157
GMM-Dif	β	não	1,4356	0,2079	0,9877	2,1720	0,4826
	β	sim	0,9999	0,0222	0,9330	1,0651	0,0222
GMM-Sys	β	não	1,3611	0,1497	1,0356	1,8523	0,3908
	β	sim	0,9999	0,0157	0,9496	1,0642	0,0156

Note. Regressions based on the system of equations (32), with population parameters: $\beta = 1$, $\rho = 0,5$, and $\phi = 0,5$. Number of simulated samples: 1,000. Param.: Estimated parameters; SD: Standard deviation; RMSE: Root Mean Squared Error. Estimators: Ordinary Least Squares (OLS); Random Effects (RE); First-differencing Generalized Method of Moments (GMM-Dif); System Generalized Method of Moments (GMM-Sys).

5 Concluding Remarks

Most of the empirical studies in corporate finance use observational data on firms with the aim of discerning causal relationships between variables by employing linear regressions. In almost all the studies in this area, however, the researcher encounters the challenge of identifying and dealing with the different endogeneity problems of the regressors, which, if ignored, may lead to inappropriate inferences. In this study, we discuss the main causes of the problem and their possible solutions, in particular when the researcher has panel data, but does not have instrumental variables that are external to the model or quasi-experimental contexts. By means of the simulated samples, which emulate some of the main characteristics of the data used in corporate finance, we illustrate the potential impact of the various sources of endogeneity, as well as some of the solutions available, comparing the relative effectiveness of different estimation methods.

The results shed light on the potential bias in the estimated coefficients when the problems of omitted variables, measurement errors of the regressors, simultaneous determination of explanatory and explained variables, or of feedback effects, also known as dynamic

endogeneity, are not adequately addressed. The joint and separate implications of these issues are illustrated using a general model and seven specific cases estimated using the Ordinary Least Squares (OLS), Fixed Effects (FE), Random Effects (RE), First-differencing GMM (GMM-Dif), and System GMM (GMM-Sys) procedures.

Our analyses show that the traditional OLS, RE, and FE estimators may be inconsistent in the presence of endogeneity problems that are quite plausible in the context of corporate finance. On the other hand, the estimation methods for panel data based on GMM that use assumptions of sequential exogeneity of the regressors present alternatives that are capable of effectively overcoming all the problems listed (provided these assumptions are valid) even if the researcher does not have good instrumental variables that are external to the model. In particular, the simulation analyses suggest that the GMM-Sys estimator (Blundell & Bond, 1998) may be an interesting option (combining low bias and high efficiency) for empirically modeling causal relationships between corporate variables.

Naturally, the effectiveness of these procedures will depend on the validity of the aforementioned assumptions of sequential exogeneity and on the adequate specification of the empirical model, something that cannot be

ensured *a priori*. For example, the validity of the sequential exogeneity assumptions depends on the absence of autocorrelation in the model errors; the assumption of temporal stability of the correlation between the regressors and the unobserved heterogeneity (see equation (18)), which is crucial for the GMM-Sys estimator, may not be realistic; if the number of instruments derived from the sequential exogeneity assumptions is very high in comparison with the number of observations of the sample, the GMM estimators analyzed here may be biased (a phenomenon known as overfitting. See Bun & Sarafidis, (2015)); Dang et al. (2015) highlight that certain characteristics of the data may generate distortions in the estimation, for example if there is censoring in the dependent variable (e.g., restricting the values of the dependent variables to the interval $[0,1]$); and there may be relevant finite sample bias when the instruments used are insufficiently correlated with the instrumented regressors (i.e., the instruments are weak. See Windmeijer, 2018; Wintoki et al., 2012).

Various diagnostic tests are available for the estimators analyzed in this study (especially those based on GMM), aiming to check the plausibility of the assumptions that ensure the validity of the estimation procedure, for example error autocorrelation tests and tests of the validity of the instrumental variables. Although they serve as important guides for the empirical modeling, unfortunately their use does not guarantee that all the problems will be detected. Wintoki et al. (2012) highlight, for example, that tests of the validity of the instruments (i.e., over-identification tests) often fail to detect endogeneity. Another problem that has been hard to diagnose, until recently, is the bias caused by weak instruments in the context of estimating dynamic panel models using GMM. Fortunately, new procedures are available that facilitate this important diagnosis (see Windmeijer, 2018). Finally, we note that the estimators used in this exercise may be adapted in order to produce appropriate inferences even in the presence of certain specification problems, for

example autocorrelation in the model errors (e.g., a few bias correction procedures are proposed and discussed by Zhou et al., 2014).

Notas

- 1 It is important to observe that the jargon used by different authors to describe these methods may vary significantly. The terminology used here follows that adopted by Arellano (2003) and Wooldridge (2010).
- 2 Another difference is that the traditional FE methods do not enable the inclusion among the regressors of time-invariant variables, unlike the RE procedures.
- 3 The validity of certain sequential exogeneity assumptions will depend on the autocorrelation pattern (or on the absence of autocorrelation) of the u_{it} errors, which can be formally tested by the researcher.
- 4 This modeling also implicitly assumes that the researcher would include any variables W_{1it}, \dots, W_{kit} available (that is, observed) and simultaneously correlated with y and with x , as discussed in section 2.1.1.

References

- Angrist, J. D., & Pischke, J. (2008). *Mostly harmless econometrics: An empiricist's companion*. New Jersey: Princeton University Press.
- Arellano, M. (2003). *Panel data econometrics*. New York: Oxford.
- 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(194), 277–297.
- Bertrand, M., & Schoar, A. (2003). Managing with style: The effect of managers on firm policies. *Quarterly Journal of Economics*, 118(4), 1169–1208.
- Black, B., & Kim, W. (2012). The effect of board structure on firm value: A multiple identification strategies approach using Korean data. *Journal of Financial Economics*, 104(1), 203–226.
- Blundell, R., & Bond, S. R. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1), 115–143.

- Blundell, R., & Bond, S. R. (2000). GMM estimation with persistent panel data: an application to production functions. *Econometric Reviews*, 19(3), 321–340.
- Bond, S. R. (2002). Dynamic panel data models: A guide to micro data methods and practice. *Portuguese Economic Journal*, 1(2), 141–162.
- Bun, M. J., & Sarafidis, V. (2015). Dynamic panel data models. In B. H. Baltagi, editor *The Oxford handbook of panel data*, (pp. 76–110). Oxford: Oxford University Press.
- Dang, V. A., Kim, M., & Shin, Y. (2015). In search of robust methods for dynamic panel data models in empirical corporate finance. *Journal of Banking & Finance*, 53, 84–98.
- Fama, E. F. & French, K. R. (2002). Testing trade-off and pecking order predictions about dividends and debt. *Review of Financial Studies*, 15(1), 1–33.
- Flannery, M. J., & Hankins, K. W. (2013). Estimating dynamic panel models in corporate finance. *Journal of Corporate Finance*, 19, 1–19.
- Frank, M. Z. & Goyal, V. K. (2003). Testing the pecking order theory of capital structure. *Journal of Financial Economics*, 67(2), 217–248.
- Glen, J., Lee, K., & Singh, A. (2001). Persistence of profitability and competition in emerging markets. *Economics Letters*, 72(2), 247–253.
- Greene, W. H. (2000). *Econometric analysis (4th ed.)*. New Jersey: Prentice Hall, Upper Saddle River.
- Himmelberg, C. P., Hubbard, R. G., & Palia, D. (1999). Understanding the determinants of managerial ownership and the link between ownership and performance. *Journal of Financial Economics*, 53(3), 353–384.
- Larcker, D. F., & Rusticus, T. O. (2010). On the use of instrumental variables in accounting research. *Journal of Accounting and Economics*, 49(3), 186–205.
- McConnell, J. J., & Servaes, H. (1995). Equity ownership and the two faces of debt. *Journal of Financial Economics*, 39(1), 131–157.
- Roberts, M. R., & Whited, T. M. (2013). Endogeneity in empirical corporate finance. In G. M. Constantinides, M. Harris, & R. M. Stulz (Eds.). *Handbook of the Economics of Finance*, (pp. 493–572). New York: Elsevier.
- Stulz, R. M. (1990). Managerial discretion and optimal financing policies. *Journal of Financial Economics*, 26(1), 3–27.
- Waring, G. F. (1996). Industry differences in the persistence of firm-specific returns. *The American Economic Review*, 86(5), 1253–1265.
- Windmeijer, F. (2018). Testing over-and underidentification in linear models, with applications to dynamic panel data and asset-pricing models (*Working Paper No. 18/696*). Department of Economics, University of Bristol, UK.
- Wintoki, M. B., Linck, J. S., & Netter, J. M. (2012). Endogeneity and the dynamics of internal corporate governance. *Journal of Financial Economics*, 105(3), 581–606.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data (2nd ed.)*. Cambridge, Massachusetts: MIT Press.
- Zhou, Q., Faff, R., & Alpert, K. (2014). Bias correction in the estimation of dynamic panel models in corporate finance. *Journal of Corporate Finance*, 25, 494–513.

Supporting Agencies: This study was financed in part by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), grant number 401562/2006-9.

Authors:

1. Lucas A. B. C. Barros, Ph.D in Finance at the University of Pennsylvania, Philadelphia, USA.

E-mail: lucasbarros@usp.br

ORCID

 0000-0002-9319-1306

2. F. Henrique Castro, Ph.D in Finance at the Columbia University in the City of New York, New York, USA. E-mail: castrohfg@gmail.com

ORCID

 0000-0001-7456-2354

3. Alexandre Di Miceli da Silveira, Ph.D in Management at Louvain-la-Neuve and Cornell Universities, Louvain, Belgium. E-mail: alexandre.miceli@fecap.br

ORCID

 0000-0002-4864-9429

4. Daniel Reed Bergmann, Doctor in Finance at the University of Sao Paulo, São Paulo, Brazil.

E-mail: danielrb@usp.br

ORCID

 0000-0003-1567-0229

Contribution of each author

Contribution	Lucas Barros	F. Henrique Castro	Alexandre Di Miceli	Daniel Reed Bergman
1. Definition of research problem	√	√	√	√
2. Development of hypotheses or research questions (empirical studies)	√	√	√	√
3. Development of theoretical propositions (theoretical Work)				
4. Theoretical foundation/ Literature review	√	√	√	√
5. Definition of methodological procedures	√	√	√	√
6. Data collection				
7. Statistical analysis	√	√	√	√
8. Analysis and interpretation of data	√	√	√	√
9. Critical revision of the manuscript	√	√	√	√
10. Manuscript Writing	√	√	√	√
11. Other (please specify which)				