Journal of Corporate Finance 19 (2013) 1–19

Contents lists available at SciVerse ScienceDirect

Journal of Corporate Finance

journal homepage: www.elsevier.com/locate/jcorpfin

Estimating dynamic panel models in corporate finance

Mark J. Flannery [a], Kristine Watson Hankins [b,*]

[a] University of Florida, Warrington School of Business, P.O. Box 117168, Gainesville, FL 32611-7168, United States
[b] University of Kentucky, 445K Gatton College, Lexington, KY 40506, United States

ARTICLE INFO

Article history:
Received 2 February 2012
Received in revised form 12 September 2012
Accepted 17 September 2012
Available online 24 September 2012

JEL classification:
G30
C23

Keywords:
Dynamic panels
Corporate finance
Econometrics

ABSTRACT

Dynamic panel models play a natural role in several important areas of corporate finance, but the combination of fixed effects and lagged dependent variables introduces serious econometric bias. Several methods of counteracting these biases are available and these methodologies have been tested on small datasets with independent, normally-distributed explanatory variables. However, no one has evaluated the methods' performance with corporate finance data, in which the dependent variable may be clustered or censored and independent variables may be missing, correlated with one another, or endogenous. We find that the data's properties substantially affect the estimators' performances. We provide evidence about the impact of various data set characteristics on the estimators, so that researchers can determine the best approach for their datasets.

1. Introduction

Dynamic panel models play an increasingly prominent role in corporate finance research. Empirically understanding payout policy, capital structure, or investment decisions arguably requires the use of firm fixed effects to control for unobserved, time-invariant differences across firms.[1] Yet uncorrected coefficient estimates for a dynamic panel model can be severely biased. Following on the observations of Nerlove (1967), Nickell (1981) established that OLS estimates of the lagged dependent variable's coefficient in a dynamic panel model are biased due to the correlation between the fixed effects and the lagged dependent variable (see also Baltagi, 2008). The bias is inversely related to panel length ("*T*"), but potentially severe biases remain even with $T = 30$ (Judson and Owen (1999)). Compustat firms have a mean (median) of 15 (11) years of annual data, well short of the number of observations required to make the bias negligible. Even when the researcher's primary concern lies elsewhere, a biased coefficient on the lagged dependent variable renders the other coefficient estimates suspect. As such, the short panel bias is a significant concern, and questions requiring dynamic panel models constitute some of the most contentious and unresolved areas of financial research.

The potential importance of choosing an appropriate estimation method for a dynamic panel model can be illustrated by recent efforts to estimate dynamic panel models of corporate leverage. Welch (2004) concludes that firms do not adjust toward target leverage; Fama and French (2002) estimate that firms adjust between 7 and 18% each year; Lemmon et al., (2008) estimate about 25% annually; Huang and Ritter (2009) estimate 17–23%; Flannery and Rangan (2006) estimate an adjustment speed above 30%. The econometric uncertainties associated with dynamic panel data have made it difficult to achieve consensus on the importance of adjustment behavior and of the factors affecting target leverage ratios. Similar problems exist in other areas of

* Corresponding author.
E-mail addresses: flannery@ufl.edu (M.J. Flannery), kristine.hankins@uky.edu (K.W. Hankins).

[1] For example, see Andres et al. (2009), Bond and Meghir (1994), Dittmar and Duchin (2010), Lemmon et al. (2008), Loudermilk (2007), Machin and Reenen (1993), and Ozkan (2000).

0929-1199/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jcorpfin.2012.09.004

corporate finance research such as corporate governance (Wintoki et al., (2012)), cash management (Dittmar and Duchin (2010)), and investment financing (Bond et al., 2003), as well as growth and banking research.

Of course, econometric techniques have evolved to correct these biases, including instrumental variables (IV), generalized method of moments (GMM) estimators, long differencing (LD), and bias correction formulae. These methods have been tested on small datasets, most of which have at most one, normally distributed independent variable. Yet corporate finance studies include multiple independent variables, of which many exhibit endogeneity and serial correlation. Therefore, an estimator's performance in simple Monte Carlo simulations may not apply to more complex empirical analysis. We examine the statistical properties of seven alternative methods for estimating dynamic panel models: OLS, standard fixed effects (FE) estimation, Arellano and Bond's (1991) difference GMM, Blundell and Bond (1998) system GMM, two variations of long differencing (Hahn et al., 2007, Huang and Ritter, 2009), and corrected least-squares (Kiviet, 1995, Bruno, 2005). By simulating data that resembles "real" corporate finance data, we evaluate the performance of these estimators under conditions that are likely to apply to corporate finance research topics. Our goal is to provide sufficient analysis that corporate empiricists can identify the estimation technique most appropriate to their data.

The paper is organized as follows. Section 2 illustrates the importance of dynamic panel estimation biases in the context of firms' capital structure choices. Section 3 explains the econometric issues and describes existing methods for addressing them. Like Petersen (2009), we utilize Monte Carlo simulations to assess the performance of various estimators in different situations. Section 4 describes how we simulate datasets. In addition to a straightforward set of independent and identically distributed (iid) simulated datasets, we also simulate datasets using a variance–covariance structure generated from actual Compustat variables. Section 5 presents some initial results. First, we confirm that most of the proposed estimation methods yield reasonably accurate coefficient estimates when data and regression residuals are generated from iid error distributions. When the simulated explanatory variables mimic Compustat data, the estimated coefficient on the lagged dependent variable remains reliable, but estimation errors for some of the other explanatory variables' coefficients increase dramatically. In Section 6, we evaluate each estimator in the presence of common corporate data features, such as missing observations, unbalanced panel lengths, and dependent variable censoring. One estimation method (Kiviet's (1995) corrected least squares dependent variable, or LSDVC) emerges as the most accurate methodology across all these dataset conditions and Blundell and Bond's (1998) system GMM estimator is often the second best choice.

Unfortunately, both have limitations. Not only does LSDVC's computer memory requirement make it difficult to apply in large datasets, it assumes exogenous regressors. Blundell Bond GMM (like Arellano Bond GMM) assumes an absence of second order serial correlation. In Section 7, we explore how violating these assumptions affects the performance of all seven estimators. Both endogenous regressors and second-order serial correlation seriously compromise many of the estimation methodologies, consistent with the theoretical literature. Perhaps surprisingly, these complications can be large enough that there are occasions when the much maligned fixed effects estimator performs best. Section 8 concludes by offering guidance about the best way to approach dynamic panel estimation in a corporate finance context.

2. An example of short panel bias

This section uses a partial adjustment model of capital structure to demonstrate the severity of the short panel bias and to illustrate the need for appropriate econometric procedures. Fischer et al. (1989) argue that adjustment costs prevent firms from adjusting completely to their optimal leverage each period. An appropriate regression specification therefore must include a lagged dependent variable to control for the prior period's capital structure. At the same time, the available data do not necessarily capture all relevant firm characteristics, perhaps including managerial risk aversion, the firm's governance structure, or cash-flow characteristics. MacKay and Phillips (2005) and Lemmon et al., (2008) conclude that fixed effects must be used to control for unobservable, time-invariant features of the firm. Yet the combination of a lagged dependent variable and firm fixed effects introduces a bias which can be substantial with short panels.

To illustrate this "short panel bias," assume that a firm's capital structure adjusts according to

$$MDR_{i,t+1} - MDR_{i,t} = \lambda\left(MDR_{i,t+1}{}^{*} - MDR_{i,t}\right) + \delta_{i,t+1} \tag{1}$$

where MDR is the ith firm's market debt ratio: the ratio of interest bearing debt to the sum of interest bearing debt plus the market value of equity,

MDR^{*} is the firm's target debt ratio,
λ is the adjustment speed toward the target, and
δ is the error term.

If target leverage depends linearly on a set of observed and unobserved firm characteristics, we can write

$$MDR_i{}^{*} = \beta X_i + F_i$$

where X_i is a vector of observable firm-specific determinants of the target MDR, β is a vector of coefficients, and F_i is a firm fixed effect. Substituting this expression for MDR^{*} into Eq. (1) yields

$$MDR_{i,t+1} = (\lambda\beta)X_{i,t} + \lambda F_i + (1-\lambda)MDR_{i,t} + \delta_{i,t+1}. \tag{2}$$

Following Flannery and Rangan (2006), *X* consists of earnings before interest and taxes scaled by total assets, market to book, depreciation scaled by total assets, the natural log of total assets (deflated to 1983 dollars), fixed assets (net PPE) scaled by total assets, an indicator for positive research and development (R&D) expenses, R&D expense scaled by total assets, and the industry median debt ratio for the firm's Fama and French (1997) industry.

Although we do not know the "true" coefficient values in Eq. (2), we can assess the sensitivity of estimated values to variations in the length of panels. The sample consists of firms with at least 30 years of continuous data from the annual CRSP/Compustat database during the period 1998–2004.[2] Following standard practice, we exclude financial firms (SIC codes 6000–6999), regulated utilities (SIC codes 4900–4999), firms with undefined Compustat formats (format codes 4, 6), and foreign firms (format code 5). We omit firm-years with a negative book value of equity or missing data for long-term debt, debt in current liabilities, or any of the leverage determinants, *X*. Size, measured as the log of total assets, is the one variable that is not a ratio and it is deflated to 1983 dollars with the consumer price index from the Bureau of Labor Statistics. To minimize the potential impact of outliers, the dependent variables and regressors are all winsorized at the 1st and 99th percentiles. The final sample includes 19,140 firm years from 638 firms, each with 30 years of data. We retain the first 30 observations for each firm.

We estimate regression Eq. (2) using three alternative methodologies: OLS (which ignores fixed effects entirely), a standard fixed effects model (which ignores the short panel bias), and Blundell and Bond's (1998) system GMM method (BB). Panel A of Table 1 presents estimation results for the 638 firms with 30 years of data. In Panels B and C, each firm's data are subdivided into multiple "imaginary" firms. For each Compustat firm in Panel A, we create three artificial "firms" in Panel B using the real firm's data from years 1–10, 11–20, and 21–30 respectively. Panel C carries this process further, creating six "firms" (years 1–5, 6–10, etc.) from each long-lived Compustat firm. Except for a "firm ID" variable that determines panel length, the data are identical across the three panels in Table 1. We thus can assess how panel length affects the various estimations when the underlying explanatory variables are otherwise identical.

Panel A shows that the estimated adjustment speed varies quite a lot across the different econometric techniques, from a low of 13% (≈1–.871) for OLS to a high of 25% (≈1–.752) for FE. With Panel B's shorter panel length ($T=10$) the OLS estimate is unchanged (as expected, since OLS ignores the unobserved firm heterogeneity that introduces a short panel bias in dynamic data). With the inclusion of firm fixed effects (FE), the estimate rises from 25% (≈1–.752 in Panel A) to 44% (≈1–.559). Meanwhile, the BB estimate rises only slightly. Panel C shortens the panel length still further, to $T=5$. The OLS estimates remain unchanged and the BB estimated adjustment speed remains at 18%, but the FE estimate again rises sharply. With 5-year panels, the alternative estimates imply adjustment speeds in the range of 13%–66% per year. While the OLS estimates are unaffected by the short panel issue, they suffer an omitted variable bias from ignoring the firm-level unobserved heterogeneity. Only the BB estimates both control for unobserved firm effects and are robust to panel length, varying between 15.4% and 18%.

Biased adjustment speed estimates also may affect the coefficients measuring the impact of various firm characteristics on desired leverage. For example, the estimated coefficient on earnings (*EBIT_TA*) is negative in all columns of Table 1, but its statistical significance fades for the BB method as *T* becomes shorter. Likewise, when $T<30$, the firm's market-to-book ratio for assets (*MB*) is significant only for the (biased) OLS estimator. In addition, the coefficient on *LnTA* is statistically significant only with the FE estimator, and the *R&D_TA* coefficient is significant only in the OLS estimation.

In sum, at least for this leverage model, estimation methodology significantly affects the inferences drawn from dynamic panel analysis. The BB method yields a consistent adjustment speed estimate, but the significance of independent variable coefficients varies with panel length. Is this result the best available, or might other statistical techniques yield more reliable estimates for dynamic corporate finance panels? We begin by expositing the econometric problem.

3. The econometrics of dynamic panel models

We can generalize Eq. (2) to write a dynamic panel model as

$$Y_{it} = \gamma Y_{it-1} + \beta X_{it} + F_i + \varepsilon_{it}. \tag{3}$$

Estimating Eq. (3) via OLS yields biased and inconsistent results because OLS omits the fixed effect, F_i. The least squares dummy variable (or "fixed effect", FE) estimate controls for the unobserved (time-invariant) heterogeneity, but it also yields biased coefficient estimates. Since Y_{it} is a function of the fixed effect, the lagged dependent variable is correlated with the error term (Baltagi (2008)). The within transformation removes the time-invariant unobserved heterogeneity from the model:

$$Y_{it}-\overline{Y_i} = \gamma\left(Y_{it-1}-\overline{Y_{i,-1}}\right) + \beta(X_{it}-\overline{X_i}) + (F_i-\overline{F_i}) + (\varepsilon_{it}-\overline{\varepsilon_i}) \tag{4}$$

but introduces a correlation between the transformed lag $\left(Y_{it-1}-\overline{Y_{i,-1}}\right)$ and the transformed error $(\varepsilon_{it}-\overline{\varepsilon_i})$ because the average error $\left(\overline{\varepsilon_i}=\sum_{i=1}^{T}\varepsilon_{it}\right)$ includes ε_{it-1}. The estimated γ therefore remains biased. (Some of the GMM estimation methods for dynamic panels first-difference Eq. (3) to eliminate the fixed effects, but the (differenced) lagged dependent variable remains correlated

[2] Firms that survive three decades clearly do not constitute a random sample of all firms, but we ignore the economic aspects of this selection bias in order to illustrate the sensitivity of estimated coefficients to alternative estimation methods. We thank Kit Baum of Boston College for this suggestion. Huang and Ritter's (2009) Fig. 5 similarly demonstrates the potential biases associated with alternative methods for estimating dynamic panel models of firm leverage.

Table 1
Panel length sensitivity. The panel length bias of various econometric specifications is compared for firms surviving at least 30 years. Panel A includes the first 30 years of data for each firm. Panel B (C) divides each firm's data into ten (five) year subgroups. OLS presents the ordinary least squares regression estimates adjusted for clustering of errors at the firm level, FE presents the fixed effects estimates (sometimes called "least squares dependent variable" or LSDV), and BB presents the Blundell Bond estimates. Year dummies are included in each specification. *MDR* is the market debt ratio. *EBIT_TA* is income before extraordinary items plus interest expense plus total income taxes, scaled by total assets. *MB* is the ratio of market to book value. *DEP_TA* is depreciation and amortization, scaled by total assets. *LnTA* is the natural log of total assets, deflated by the consumer price index to 1983 dollars. *FA_TA* is net PPE, scaled by total assets. *R&D_Dum* is an indicator equaling one if research and development expense is positive, else zero. *R&D_TA* is research and development expense, scaled by total assets. *Ind_Median* is the median debt ratio for the firm's Fama and French (1997) industry. P-values are listed in parentheses.

$$MDR_{i,t+1} = (\lambda\beta)X_{i,t} + \lambda F_i + (1-\lambda)MDR_{i,t} + \delta_{i,t+1}. \quad (2)$$

	Panel A: Full 30 years			Panel B: Ten year subgroups			Panel C: Five year subgroups		
	OLS	FE	BB	OLS	FE	BB	OLS	FE	BB
MDR	0.871	0.752	0.846	0.871	0.559	0.825	0.871	0.337	0.820
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
EBIT_TA	−0.033	−0.042	−0.113	−0.033	−0.071	−0.090	−0.033	−0.079	−0.089
	(0.004)	(0.000)	(0.003)	(0.004)	(0.000)	(0.047)	(0.004)	(0.000)	(0.173)
MB	−0.005	−0.002	−0.004	−0.005	−0.001	0.000	−0.005	−0.001	0.003
	(0.000)	(0.036)	(0.078)	(0.000)	(0.555)	(0.929)	(0.000)	(0.718)	(0.722)
DEP_TA	−0.308	−0.501	−2.243	−0.308	−0.497	−2.248	−0.308	−0.499	−1.818
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.082)
Ln(TA)	0.001	0.017	−0.001	0.001	0.046	0.001	0.001	0.068	0.026
	(0.128)	(0.000)	(0.761)	(0.128)	(0.000)	(0.854)	(0.128)	(0.000)	(0.174)
FA_TA	0.038	0.059	0.173	0.038	0.094	0.416	0.038	0.129	−0.073
	(0.000)	(0.000)	(0.134)	(0.000)	(0.000)	(0.035)	(0.000)	(0.000)	(0.839)
R&D_Dummy	0.001	−0.001	0.033	0.001	−0.002	0.139	0.001	−0.004	0.282
	(0.417)	(0.797)	(0.511)	(0.417)	(0.612)	(0.085)	(0.417)	(0.340)	(0.034)
R&D_TA	−0.049	0.004	0.084	−0.049	0.016	−0.360	−0.049	0.089	−0.641
	(0.094)	(0.932)	(0.808)	(0.094)	(0.775)	(0.581)	(0.094)	(0.266)	(0.590)
Ind_Median	0.008	−0.004	−0.385	0.008	0.010	−0.157	0.008	−0.035	0.437
	(0.326)	(0.737)	(0.057)	(0.326)	(0.514)	(0.684)	(0.326)	(0.076)	(0.489)
Constant	0.016	−0.272	0.139	0.016	−0.804	−0.135	0.016	−1.180	−0.682
	(0.098)	(0.000)	(0.302)	(0.098)	(0.000)	(0.490)	(0.098)	(0.000)	(0.168)
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# Obs	19,140	19,140	19,140	19,140	19,140	19,140	19,140	19,140	19,140
# Groups		638	638		1914	1914		3828	3828
Adj speed $(1-\lambda)$	13%	25%	15%	13%	44%	18%	13%	66%	18%

with the differenced residual (ε)). The bias declines with panel length because ε_{it-1} becomes a smaller component of the average error term as *T* increases. In other words, with higher T the correlation between the lagged dependent variable and the regression errors becomes smaller.

Econometric techniques have been derived to correct this short panel bias. A traditional instrumental variables (IV) approach offers one good option, provided one can identify reliable instruments. Arellano and Bond (1991) use a generalized method of moments framework to develop valid instruments. They first-difference the panel data to remove the time-invariant fixed effect and show that the lagged dependent variables' values (levels) constitute legitimate instruments for the first-differenced variable, provided that the residuals are free from second-order serial correlation. Arellano and Bond's (relatively limited) Monte Carlo simulations show that their AB (or "difference GMM") method outperforms OLS and fixed effects (FE) estimators when the regression residuals are uncorrelated. However, the lagged levels may provide little information about the first-differenced variable particularly if they are serially correlated (Arellano and Bover (1995), Blundell and Bond (1998)),[3] Blundell and Bond (1998) suggest an alternative GMM "system" estimator: in addition to the first-differencing used by Arellano Bond, Blundell and Bond utilize the lagged first differences as instruments in a non-transformed (levels) equation. Both Arellano Bond and Blundell Bond can handle endogenous regressors, using the lagged levels or first differences of those variables as instruments.

In part because the AB and BB instruments are invalidated by second-order autocorrelation (Baltagi (2008)), Hahn et al. (2007) derived a "long difference" instrumental variable estimation technique. Assuming balanced panels, Hahn et al. show that combining multi-period differencing with longer lagged instrument choices can produce less biased estimates than the AB or BB approaches. This derivation assumes balanced panels: each sample firm has the same number of observations. Because long differencing is new theory, no existing empirical or theoretical work evaluates the performance of long differencing in unbalanced panels. In applying the long difference estimator to unbalanced Compustat panels, Huang and Ritter (2009) use the same differencing interval for all firms, regardless of their panel lengths. They present results for lags of 4, 8, 18, and 28 years. We implement two variants of the Hahn et al. concept, as explained in Appendix A. First, we utilize Huang and Ritter's 4-year differencing (LD4), whose relatively short differencing

[3] Recall that using weak instruments can result in worse estimates than those derived from un-adjusted variables (Nelson and Startz (1990), Bound et al. (1995)).

interval maximizes the number of included data points. Second, we use each sample firm's longest available differencing interval (LD). Because other ways to implement LD with unbalanced panels may be superior to the ones evaluated here, we view these results as preliminary. While long difference estimators could be adapted to handle endogenous estimators in a manner similar to Arellano Bond and Blundell Bond, Hahn et al. do not include independent regressors in their simulations and our extensions of their method assume exogenous regressors.

The techniques discussed so far have concentrated on computing valid instruments with which to remove the correlation between the transformed lagged dependent variable and the transformed error term. Kiviet (1995) takes a different approach by computing an explicit, data-dependent correction for the fixed effects bias in short panels. This bias-corrected least squares dummy variable estimator (LSDVC) removes an approximated small sample bias from the FE estimator. Judson and Owen (1999) documents that LSDVC dominates AB and BB for balanced panels of all lengths. Bruno (2005) computes the bias correction for unbalanced dynamic panels, making it possible to include firms that enter and leave the dataset during the study period. LSDVC's potential disadvantages include its assumption that the regressors are strictly exogenous. Our Monte Carlo simulations provide the first information (to our knowledge) about how LSDVC performs with multiple independent variables, missing or censored data, endogeneity, and serially correlated errors.

We evaluate and compare seven econometric methods for estimating dynamic panel models on datasets with a variety of known characteristics. Table 2 summarizes the underlying assumptions of each estimation methodology.

OLS Ordinary Least Squares ignores the data's panel structure and generally produces an upward-biased coefficient estimate for the lagged dependent variable in the presence of unobserved heterogeneity (Bond (2002)). We use the Stata procedure "reg" to generate these coefficient estimates.

FE Fixed Effects incorporates the data's panel structure but ignores the correlation between the lagged dependent variable and the regression error. FE yields a downward-biased coefficient estimate for the lagged dependent variable (Nickell (1981)). We use the Stata procedure "xtreg, fe" to generate these estimates.

The following five "advanced" estimators use alternative techniques to avoid the biases associated with FE.

AB Arellano and Bond's (1991) difference GMM first-differences the linear regression model and uses lagged dependent variable levels to instrument for the first difference of the lag. We employ the Stata procedure 'xtabond', which defaults to using one lag of the exogenous variables as the instrument set.

BB Blundell and Bond's (1998) system GMM estimates a two-equation system of the regression in levels and in first differences. We use the Stata procedure "xtdpdsys" and limit the maximum number of lags to two.[4] We specify the explanatory variables as predetermined,[5] not fully exogenous.

LD4 Four Period Long Differencing replicates the Huang and Ritter (2009) implementation of Hahn et al.'s (2007) estimator. They consider several differencing windows, but we implement their 4-period variation in our simulations because it excludes the fewest firms (see Appendix A). Our estimates are computed with Stata code kindly provided by Rongbing Huang.

LD Longest Differencing is an alternative adaptation of Hahn et al.'s (2007) balanced panel estimator, in this case allowing for unbalanced panels. No literature describes how to implement an LD estimator with unbalanced panels; our procedure is described in Appendix A.

LSDVC Least Squares Dummy Variable Correction corrects the biased FE-estimated coefficients, using an estimate of the short-panel bias computed from each firm's data. It assumes that the independent variables are exogenous. Our estimates are computed with the user-written Stata procedure, 'xtlsdvc' (Bruno (2005)), which requires a vector of coefficient starting values. We experimented with using either the AB or the BB initial estimates, but found that the LSDVC estimates are robust to the initial matrix selection. We therefore report results derived from AB estimates of the initial coefficient matrix.

The econometrics literature contains Monte Carlo tests of each of these seven methods' efficacy and accuracy, but most of those studies focus on datasets with at most one exogenous variable and limited variation in the panels' size. For example, Kiviet (1995) evaluates LSDVC using a balanced panel of one hundred units with $T = 3$ or 6. Corporate finance data sets generally feature a larger matrix, unbalanced panels, and explanatory variables that might be endogenous or serially correlated. Moreover, the GMM methods' instruments are theoretically invalid if the regression incorporates second-order serial correlation, but we know little about the severity of the resulting biases.

This paper exposes seven estimation methods to a variety of dataset features. We compare estimators on the basis of their root mean squared errors (RMSEs): the square root of the average squared coefficient estimation error. RMSE provides a common benchmark for comparing dynamic panel estimators (Kiviet (1995), Blundell and Bond (1998), Wooldridge (2009)). We pay attention to the lagged dependent variable's RMSE, and to the average RMSE on the other estimated coefficients (Section 5.2 discusses some results for individual explanatory variables). In unreported results, our conclusions are unchanged when we evaluate the estimators' mean absolute errors (MAE).

[4] We also tested a one-step Blundell Bond GMM estimator, which yielded larger biases than the reported two-step BB results. Those results remain untabulated.

[5] When endogenous variables are introduced in Section 5, we run the Stata xtdpdsys procedure both with all variables specified as predetermined as well as with the three endogenous variables specified as endogenous. Since the RMSE errors are similar, if not larger, with the second specification, we continue to report all Blundell Bond results using the predetermined option.

Table 2
Methodologies' assumptions. Each of the methodologies we examine is designed to cope with specific dataset features, such as "Unobserved heterogeneity", "Dynamic panel data", etc. The columns in this table indicate whether each methodology is designed to produce appropriate estimates under the indicated condition.

	Unobserved heterogeneity	Dynamic panel data	Second order serial correlation	Unbalanced panel data	Endogenous variables
1. OLS	No	No	Yes	Yes	No
2. Fixed effects	Yes	No	Yes	Yes	No
3. Arellano Bond	Yes	Yes	No	Yes	Yes
4. Blundell Bond	Yes	Yes	No	Yes	Yes
5. Longest diff	Yes	Yes	Yes	Assumed, but Untested	Assumed, but Untested
6. 4 Period diff	Yes	Yes	Yes	Assumed, but Untested	Assumed, but Untested
7. LSDVC	Yes	Yes	Yes	Yes	No

4. Constructing simulated data

The true model underlying our generated data takes the general form:

$$y_{it} = \gamma y_{it-1} + \sum_{j=1}^{J} \beta_j x_{jit} + \eta_i + \varepsilon_{it} \tag{5}$$

where γ and β_j are the primary coefficients to be estimated, x_{jit} are explanatory variables, η_i is a firm fixed effect, and ε_{it} is the residual. Unlike most prior Monte Carlo studies, we include multiple ($j>1$) explanatory variables and one of our main concerns is the impact of the explanatory variables' characteristics on estimators for Eq. (5). The explanatory variables evolve according to

$$x_{ijt} = \rho_j x_{ijt-1} + \alpha_1 y_{it-1} + \alpha_2 \eta_i + \xi_{ijt}, \tag{6}$$

in which ρ_j, α_1, and α_2 are parameters that vary across generated data sets and ξ_{ijt} is normally distributed. This general specification permits us to control the explanatory variables' serial correlations (ρ_j). We can also make some or all of the explanatory variables endogenous – that is, correlated with the lagged dependent variable (y_{t-1}) and the firm fixed effect (η_i) – by setting α_1 or $\alpha_2 \neq 0$. We treat the endogeneity issue in Section 7.1 and meanwhile generate purely exogenous independent variables according to

$$x_{ijt} = \rho_j x_{ijt-1} + \xi_{ijt}. \tag{6a}$$

Finally, we can examine the effect of serially correlated residuals on the various estimators by varying δ_1 and δ_2 in

$$\varepsilon_{it} = \delta_1 \varepsilon_{i,t-1} + \delta_2 \varepsilon_{i,t-2} + \omega_{it}. \tag{7}$$

We use various parameterizations of Eqs. (5), (6a), and (7) to generate simulated data as in Arellano and Bond (1991) and Judson and Owen (1999): start with $x_{ji0}=0$ and $y_{ji0}=0$, generate a panel of length $T+10$ for each firm and then drop the first 10 observations. We generate 500 datasets under each set of parameter assumptions.[6] For each generated dataset, we estimate the model (5) using each of the seven estimation methodologies and store the estimated $\hat{\gamma}$ and $\hat{\beta}_j$ along with their standard errors. We then evaluate the performance of the methodologies by comparing these estimates with the true coefficient values used to generate the simulated dataset. Appendix B provides a summary of the data set parameterizations for which we compare estimators.

4.1. General principles

We subsequently refer to two broad groups of simulated datasets, which differ in how we treat the explanatory variables. The "iid" datasets generate each value of x_{ijt} independently of the other variables' values. The "Compustat" datasets uses a joint normal distribution whose variance–covariance matrix is derived from actual Compustat data to generate the panel of independent regressors. The iid and the Compustat simulated datasets are constructed using the same basic procedure. Each simulated data set includes a sample of N = 500 firms. Since the panel length (T) is a primary interest, we generate panels of three lengths: $T=6$, 12, or 30.

- $T=6$ represents an estimator's performance in shorter panels.
- $T=12$ corresponds to the median panel length among Compustat firms observed annually.
- $T=30$ represents a longer Compustat panel, corresponding to Judson and Owen (1999) and to Huang and Ritter's (2009) 28-year differencing interval.

[6] Many readers will be accustomed to seeing Monte Carlo simulations with thousands of dataset "draws", or more. We limit our analysis to 500 datasets (replications) for each parameter and model variation since the user-written Stata command for LSDVC (xtlsdvc) is quite computationally intensive and time consuming. Previous Monte Carlo evaluations of dynamic panel estimators also regularly have used 500 or fewer replications.

These choices for N and T imply datasets with between 3000 and 15,000 observations.

We use a random number generator in Stata to generate firm-level fixed effects (η) that are uniformly distributed over the interval [−1, 1].[7] The panel innovations (ε) in Eq. (5) are generated using a N(0,1) distribution. We alternate γ (in Eq. (5)) between values of 0.2 and 0.8 to assess the effect of the dependent variable's serial correlation. In Eq. (5), $\beta_j = 0.2$ for all j, as in Kiviet (1995). Initially, we set $\delta_1 = \delta_2 = 0$ in Eq. (7) but we vary the values in later analysis to consider the impact of serial correlation in the residuals. Both the iid and Compustat datasets contain seven independent variables.[8]

4.2. Simulated "iid" datasets

The Monte Carlo simulations previously used in the econometrics literature to evaluate dynamic panel methodologies often have included only one, serially-uncorrelated explanatory variable in the simulated model. Because corporate finance is rarely that simple, we generate iid datasets containing multiple explanatory variables for our baseline analysis. We generate a dynamic panel based on the stochastic properties of the seven exogenous independent variables in Eq. (1). To capture within-firm persistence, we amend Eq. (6a) to include firm fixed effects (Ω_i) to control for unobserved heterogeneity but assume exogenous regressors:

$$x_{ijt} = \rho_j x_{ijt-1} + \Omega_{ij} + \xi_{ijt}, \tag{6b}$$

where Ω_{ij} is a firm-specific effect on the value of the jth independent variable at the ith firm. Serial correlation in the explanatory variables is introduced by estimating Eq. (6b) for each of the seven explanatory variables in Eq. (2) (the sample includes all CRSP/Compustat firms except utilities and financial firms from 1962 through 2006). Using $\hat{\rho}_j$ estimated from Eq. 6b, we generate exogenous variables via Eq. (6a) assuming $\xi_{ijt} \sim N(0, 1)$. These $\hat{\rho}_j$ are reported in Panel A of Table 3. Our "iid" variables thus have stationarity derived from Compustat data, but their innovations are independent of one another and all have the same innovation variance.

4.3. Simulated "Compustat" datasets

Corporate data may differ substantially from the assumptions of the iid simulated datasets, which have uncorrelated explanatory variables with identical volatilities. Yet actual firms' characteristics are likely to be correlated with one another (e.g. older firms tend to be larger and to have less growth potential (lower M/B)) and the explanatory variables' volatilities may differ. To capture a more realistic structure in the explanatory variables, we calculate a variance–covariance matrix for the seven sets of residuals from the estimation of Eq. (6b) and use the Stata command "drawnorm" to generate residuals $\xi_{ijt}, j = 1, 7$ in Eq. (6a). In other words, the seven explanatory variables are distributed multivariate normal with the computed variance–covariance matrix shown in Panel B of Table 3.

Table 3 Panel A indicates that the explanatory variables are quite highly serially correlated within firms. Furthermore, Panel B indicates the independent variables differ substantially in the variances of their innovations; Depreciation/TA, R&D/TA and Industry Median have particularly small variances (by comparison, the iid explanatory variables' residual variances are 1.0).

One could imagine other ways to simulate "realistic" corporate data, but this one seems reasonable and allows us to examine the impact of a non-iid error structure in the measured variables. Although the specific variables relate to the capital structure literature, they are common control variables in a number of corporate finance contexts and we believe that the estimation results are applicable to a broader set of corporate finance questions, such as investment, payout policy, or corporate governance.

5. Initial results

This section compares the seven estimation methods across the iid and Compustat-based simulated datasets.

5.1. Estimation results for iid simulated datasets

We start by comparing the seven estimators using iid datasets constructed with serially uncorrelated residuals (that is, $\delta_1 = \delta_2 = 0$ in Eq. (7)). Each estimator is applied to six sets of 500 simulated datasets: three (balanced) panels of length 6, 12, or 30 years, each with high or low persistence ($\gamma = 0.2$ or 0.8) for the dependent variable. The criteria for estimator performance are the RMSE of the lagged dependent variable's coefficient and the average RMSE of the seven independent variable coefficients. These results are reported in Table 4 and summarized in the left column of Fig. 1. A "better" estimator has a smaller RMSE.

We focus our discussion on the $T = 12$ results, presented in the middle columns of Table 4 and the middle panel of the left column of Fig. 1 (the other panel lengths yield qualitatively similar results). Consider first the estimated coefficient on the lagged dependent variable. When $\gamma = 0.2$, all five advanced methods yield equally accurate estimates of the lagged dependent variable's coefficient (RMSE ≈ 0.02). When $\gamma = 0.8$, however, LD4's RMSE quintuples, while the RMSEs of the other GMM methods and LSDVC remain approximately unchanged. This sensitivity of the LD4 estimates to the lagged dependent variable's value is noteworthy. LD also is less accurate for $\gamma = 0.8$ when $T = 6$ but the other advanced estimators' RMSEs are not very sensitive to the true value of γ.

[7] In unreported robustness analysis, we vary the magnitude of the fixed effect and find that this assumption materially affects only the OLS estimates.

[8] The seven independent variables are chosen to mimic Flannery and Rangan (2006). However, in unreported robustness analysis, we vary the number of independent variables and find that this simulation choice does not affect the interpretation.

Table 3
Compustat variables' properties. This table summarizes the seven Compustat variables used as independent variables. Panel A presents the variables' within firm serial correlation and Panel B presents the variance–covariance matrix for the residuals. The data is from Faulkender et al. (2012) and spans from 1965 to 2006.

	EBIT/TA	M/B	Depreciation/TA	Ln(TA)	Fixed assets/TA	R&D/TA	Industry median
Panel A: Serial correlation properties							
Serial correlation	0.455	0.259	0.085	0.844	0.254	0.197	0.690
Std error	0.003	0.003	0.003	0.003	0.003	0.002	0.002
Panel B: Variance–covariance matrix							
EBIT/TA	0.223						
M/B	−0.468	4.470					
Depreciation/TA	−0.003	0.000	0.002				
Ln(TA)	0.010	−0.154	−0.001	3.937			
Fixed assets/TA	0.001	0.008	0.000	−0.085	0.029		
R&D/TA	−0.005	0.012	0.000	−0.003	0.001	0.004	
Industry median	0.000	−0.005	0.000	−0.007	0.000	0.000	0.003

Unsurprisingly, FE and OLS estimators perform poorly. Turning to the independent variables' coefficients, all the advanced estimators have relatively low RMSE (<0.025), but LSDVC is lowest (RMSE≈0.013).[9] Note that FE's independent variable coefficients are as accurate as those of the advanced estimators, despite FE's poor estimates for the lagged variable's coefficient. The impact of longer panels (moving from top to bottom in Fig. 1) conforms to theory: estimation errors decline for all the panel estimators, including FE. Differences among the advanced estimators diminish as T rises.

To summarize, the iid results in Fig. 1 and Table 4 indicate that the advanced estimators AB, BB, and LSDVC do their job quite well when the independent variable's residuals are normally distributed. Although the RMSEs differ slightly across estimation methods, they are *all* quite small. FE and LD become more accurate with longer panels. Both long difference approaches are affected by stronger lag persistence. Overall, LSDVC appears to be the best choice.

5.2. Estimation results for Compustat-based simulated datasets

Estimation results from the Compustat-based simulated datasets are reported in Table 5 and plotted in the right half of Fig. 1. This data differs from the iid data in that the panel innovations (or errors) are drawn randomly using the variance–covariance matrix of real Compustat variables. Note that all Fig. 1 vertical axes have the same scale for ease of comparison. It is readily seen that the accuracy of lagged coefficient estimates does not change substantially when we replace iid independent variables with Compustat-like variables. Thus, the advanced estimators estimate the lagged dependent variable coefficient quite well in practice.[10] However, the mean RMSE for independent variables' coefficients increases dramatically. Instead of a RMSE<0.05, the estimates' RMSEs with "Compustat" data sometimes exceed 0.20. Higher T reduces estimation errors with Compustat-style data, but the mean explanatory variable's coefficient has a large RMSE (≈0.08) even with $T=30$, compared to 0.01 with iid-based data with the same panel length. The long differencing methodologies perform relatively poorly, while FE and OLS again do fairly well on the explanatory variables' coefficients. LSDVC dominates the other estimators for Compustat-based data given its combined performance in estimating both the lag and independent variables.

Table 6 decomposes the independent variables' average RMSEs, reporting the RMSE for each of the seven individual explanatory variables separately. Recall that the Compustat variables are not included directly in the simulations. Rather, we use the variance covariance matrix of these variables to simulate our panel dataset. Some explanatory variables (e.g. those simulated from M/B and ln(TA)) exhibit RMSEs close to, or below, those attained with the iid datasets. Other variables' coefficients (e.g. those simulated from Depreciation/TA, R&D/TA, Industry Median) have extremely high estimation errors, which appear to be inversely related to the variables' within-firm variance.[11] The rightmost column in Table 6 reports a strong negative correlation between the explanatory variables' variances and their coefficients' average RMSEs for all panel lengths ($T=6$, 12, or 30). This highlights a known limitation of FE and first differencing approaches, such as AB and BB. While Wooldridge (2009) notes the difficulty of accurately estimating mostly time invariant (or sluggish) variables, this issue is often overlooked in finance. To the extent that researchers are concerned with the economic significance of the independent variables, those with low within-firm variation may not be reliably estimated.

[9] Blundell and Bond (1998) document the poor performance of GMM-style instruments with highly persistent lags for the lagged dependent variable with smaller panels (small N). However, in our simulations of 500 firms, we find relatively small differences between $\gamma=0.2$ or 0.8 for LSDVC, AB, BB, and LD when $T=12$ or $T=30$.

[10] Several recent papers challenge the power of partial adjustment models to reject other processes that might be generating corporate data (Chang and Dasgupta (2009), Iliev and Welch (2010), Elsas and Florysiak (2011)). Our results here indicate that if the data are generated by a dynamic panel model, then the advanced estimators can estimate at least the autoregressive parameter quite accurately (we address the censored dependent variable issue in Section 6.3).

[11] Gormley and Matsa (2012) show that the inclusion of the industry median, a common practice in capital structure research, introduces measurement error bias. However, we simply use the moment properties of the industry median and thus do not face the omitted variable concern.

Table 4
Balanced panels with iid innovations. For each of the six base parameter variations ($T=6$, 12, or 30, $\gamma=0.2$ or 0.8), 500 dynamic panels are generated. This table summarizes how each estimation method performs when used to estimate the model with each type of simulated data set. The "Lag RMSE" rows report the average RMSE of the estimated coefficients on the lagged dependent variable across the 500 simulations. The "Xs RMSE" rows report the average of the seven explanatory variables' estimated regressions, averaged across the 500 simulations For each method the percentage of the 500 simulations with estimated coefficients that are larger than the true coefficient are reported in *% Lag* (*Xs*) *over*. Numbers in **bold** (*italics*) are the **smallest** (*second-smallest*) RMSE for the lag or Xs, within each column.

		$T=6$		$T=12$		$T=30$	
	Lag value (γ):	0.2	0.8	0.2	0.8	0.2	0.8
OLS	Lag RMSE	0.217	0.127	0.217	0.125	0.217	0.123
	Xs RMSE	0.029	0.030	0.025	0.026	0.022	0.023
	% Lag over, % Xs over	100/20	100/23	100/17	100/21	100/10	100/14
Fixed effects	Lag RMSE	0.162	0.242	0.079	0.108	0.032	0.037
	Xs RMSE	**0.022**	**0.024**	*0.014*	*0.015*	**0.008**	*0.009*
	% Lag over, % Xs over	0/56	0/35	0/65	0/56	0/65	0/67
Arellano Bond	Lag RMSE	0.032	0.094	0.017	0.030	*0.009*	*0.010*
	Xs RMSE	0.029	*0.030*	0.018	0.018	0.011	0.011
	% Lag over, % Xs over	48/52	38/47	45/52	33/51	40/52	23/52
Blundell Bond	Lag RMSE	*0.026*	**0.041**	*0.016*	*0.026*	*0.009*	*0.010*
	Xs RMSE	0.032	0.032	0.021	0.021	0.011	0.012
	% Lag over, % Xs over	58/51	95/58	58/50	93/57	55/53	86/56
Longest difference	Lag RMSE	0.043	0.127	0.027	0.035	0.024	0.017
	Xs RMSE	0.042	0.042	0.025	0.025	0.021	0.021
	% Lag over, % Xs over	36/50	0/52	46/51	17/53	50/50	40/51
4 Period diff.	Lag RMSE	0.043	0.127	0.025	0.125	0.019	0.124
	Xs RMSE	0.042	0.042	0.017	0.019	0.010	0.012
	% Lag Over, % Xs Over	36/50	0/52	16/56	0/54	4/57	0/55
LSDVC	Lag RMSE	**0.025**	*0.064*	**0.013**	**0.016**	**0.007**	**0.005**
	Xs RMSE	*0.024*	**0.024**	**0.013**	**0.014**	**0.008**	**0.008**
	% Lag over, % Xs over	33/50	4/44	45/52	20/51	48/52	45/52

5.3. Implications

The baseline simulations highlight a few major points. The lagged dependent variable is estimated equally well for iid or Compustat explanatory variables. LSDVC generally is most accurate with respect to the independent variable coefficient estimates while BB is the next best choice. Further, Table 6 shows that the estimated coefficients are less accurate for explanatory variables with low within-firm variation. The mean RMSE for the explanatory variables' coefficients are an order of magnitude larger with Compustat-based regressors than with the iid (and innovations' variances = 1) datasets. The problem derives not from stationarity per se (since the iid variables share the same stationarity parameters), but with relatively small innovations in an independent variable. This is a serious concern as many Compustat variables commonly used in corporate finance analysis share this attribute. While FE often exhibits a low RMSE for the exogenous variables, LSDVC is the overall best option for a balanced dynamic panel on any length with exogenous Compustat regressors.

A further implication relates to the choice between annual and quarterly data. Theory predicts that longer panels are estimated with greater accuracy. But at the same time, quarterly data almost surely include smaller innovations between periods. So the potential switch to more frequent observations may increase the time-invariance of some variables and thus increase the difficulty of estimating accurate coefficients.[12]

We now compare alternative estimation methods in the presence of known data challenges.

6. Further investigations of corporate finance dataset properties

Corporate finance datasets commonly include at least some of the following features:

- Unbalanced panels
- Missing observations
- Censored or clustered data

Evaluating the effects of these data features on alternative estimators should help researchers determine which estimators to use, given what they know about their dataset properties. Given the larger RMSEs associated with Compustat-consistent data, we pay no further attention to iid datasets. We also limit these tests to panels of moderate length ($T=12$), which most closely resembles the average annual Compustat panel length.

[12] Wintoki et al. (2012) also notes the estimation concerns surrounding highly persistence regressors.

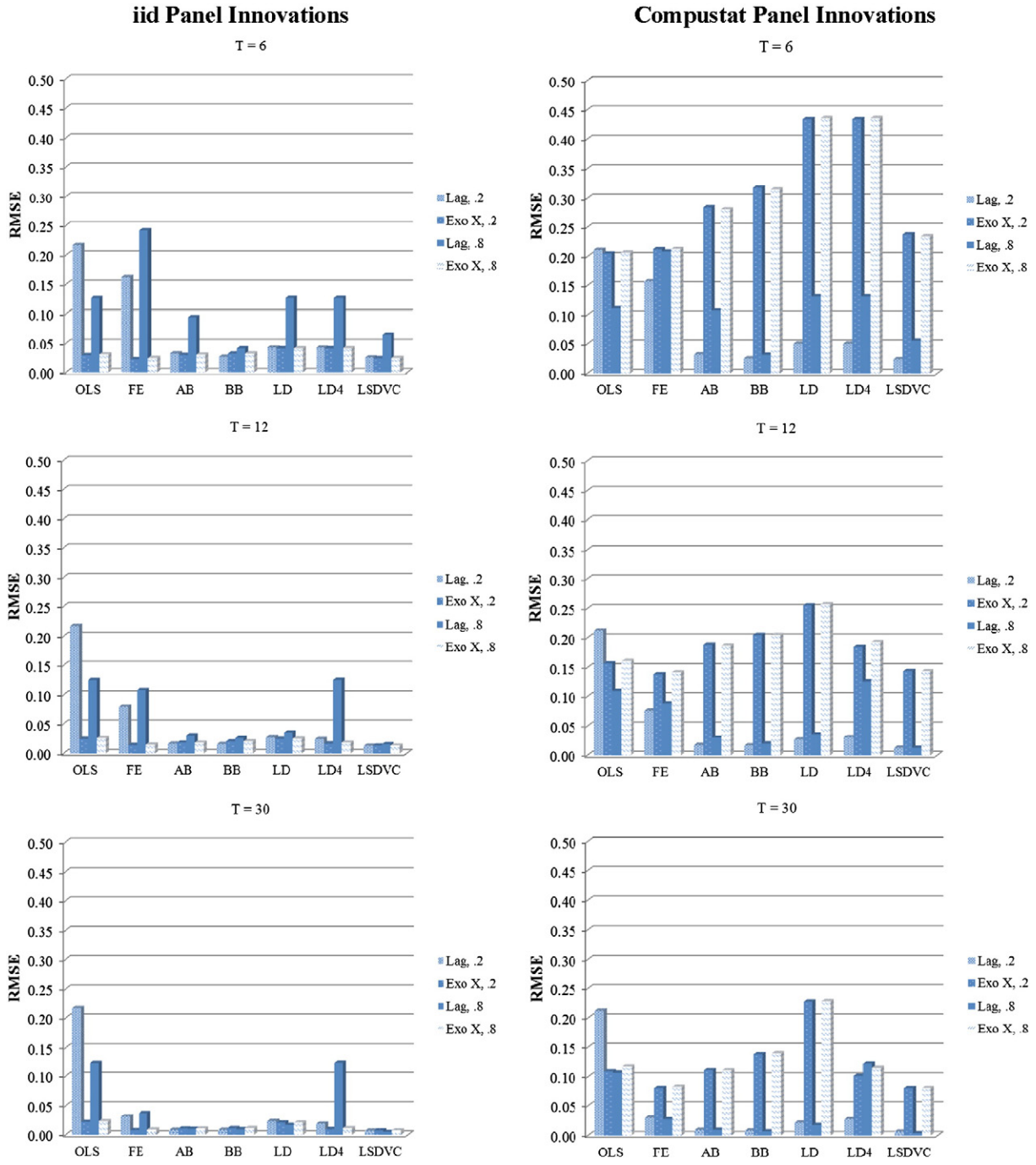

Fig. 1. Balanced panels. Source: Tables 4 and 5. The left column summarizes the balanced panel simulations with iid innovations and the right column summarizes the Compustat based simulations. For both high and low persistence ($\gamma = 0.2$ or 0.8), we present the lag dependent variable RMSE (*Lag*) as well as the average RMSE for the seven independent variables (*Exo X*).

6.1. Unbalanced panels

Thus far, our simulations have evaluated only *balanced* panels, and the same is true for most of the existing Monte Carlo evaluations of dynamic panel estimators. The GMM techniques for estimating dynamic panel coefficients utilize lagged values and lagged differences of varying degrees (from $t-2$ all the way back to $t=1$). Since panel length differences within a dataset affect the availability of lags, they may also affect the performance of such estimators. To generate unbalanced panels, we maintain a constant average $T=12$, but vary the

Table 5
Balanced panels with Compustat innovations. For each of the six base parameter variations ($T=6$, 12, or 30, $\gamma=0.2$ or 0.8), 500 dynamic panels are generated. This table summarizes how each estimation method performs when used to estimate the model with each simulation. The "Lag RMSE" rows report the average RMSE of the estimated coefficients on the lagged dependent variable across the 500 simulations. The "Xs RMSE" rows report the average of the seven explanatory variables' estimated regressions, averaged across the 500 simulations. For each method the percentage of the 500 simulations with estimated coefficients that are larger than the true coefficient are reported in *% Lag* (*Xs*) *over*. Numbers in **bold** (*italics*) are the **smallest** (*second-smallest*) RMSE for the lag or Xs, within each column.

		$T=6$		$T=12$		$T=30$	
	Lag value (γ):	0.2	0.8	0.2	0.8	0.2	0.8
OLS	Lag RMSE	0.212	0.112	0.212	0.110	0.212	0.107
	Xs RMSE	*0.206*	**0.207**	0.157	0.162	0.109	0.118
	% Lag over, % Xs over	100/34	100/33	100/33	100/33	100/31	100/32
Fixed effects	Lag RMSE	0.158	0.209	0.077	0.089	0.031	0.028
	Xs RMSE	**0.213**	0.213	**0.139**	**0.142**	**0.081**	*0.083*
	% Lag over, % Xs over	0/58	0/51	0/61	0/59	0/61	0/62
Arellano Bond	Lag RMSE	0.033	0.108	0.018	0.030	0.010	0.010
	Xs RMSE	0.284	0.280	0.188	0.187	0.111	0.111
	% Lag over, % Xs over	47/51	32/48	46/52	29/52	39/54	24/55
Blundell Bond	Lag RMSE	*0.026*	**0.032**	*0.017*	*0.020*	*0.009*	*0.007*
	Xs RMSE	0.318	0.315	0.205	0.205	0.138	0.140
	% Lag over, % Xs over	61/50	92/54	63/51	89/55	55/53	79/53
Longest difference	Lag RMSE	0.051	0.132	0.027	0.036	0.022	0.018
	Xs RMSE	0.434	0.436	0.256	0.258	0.228	0.229
	% Lag over, % Xs over	25/51	0/55	46/50	15/53	50/51	42/52
4 Period diff.	Lag RMSE	0.051	0.132	0.032	0.126	0.028	0.122
	Xs RMSE	0.434	0.436	0.185	0.193	0.102	0.115
	% Lag over, % Xs over	25/51	0/55	7/54	0/56	0/58	0/56
LSDVC	Lag RMSE	**0.024**	*0.056*	**0.013**	**0.013**	**0.007**	**0.004**
	Xs RMSE	0.238	0.235	*0.144*	*0.144*	**0.081**	**0.081**
	% Lag over, % Xs over	35/51	6/49	45/52	25/53	51/52	47/53

Table 6
Magnitude of Compustat innovation and estimation error. We report the average RMSE for each independent variable, averaged across the lag variations of 0.2 and 0.8. Recall that these variables all carry true coefficients of 0.20. Average RMSE row is calculated across all seven independent variables within each methodology and this average is presented in the remaining tables and figures as the X's RMSE. The Mean RMSE column calculates the average across the seven methodologies for each independent variable. The correlation is calculated from the independent variables' Mean RMSE and their variance.

	OLS	FE	AB	BB	LD	LD4	LSDVC	Mean RMSE	X_i's Variance	Correlation
Panel A. T=6										
EBIT/TA_{t-1}	0.044	0.053	0.071	0.078	0.102	0.102	0.057	0.073	0.223	−0.671
M/B_{t-1}	0.014	0.014	0.016	0.017	0.023	0.023	0.013	0.017	4.470	
Depreciation/TA_{t-1}	0.497	0.510	0.665	0.728	1.161	1.161	0.583	0.758	0.002	
$Ln(TA)_{t-1}$	0.049	0.016	0.016	0.017	0.021	0.021	0.012	0.022	3.937	
Fixed assets/TA_{t-1}	0.192	0.153	0.160	0.186	0.280	0.280	0.142	0.199	0.029	
R&D/TA_{t-1}	0.342	0.342	0.447	0.508	0.751	0.751	0.385	0.504	0.004	
Industry $median_{t-1}$	0.305	0.403	0.599	0.683	0.707	0.707	0.461	0.552	0.003	
Average RMSE	0.206	0.213	0.282	0.317	0.435	0.435	0.236			
Panel B. T=12										
EBIT/TA_{t-1}	0.032	0.033	0.047	0.054	0.063	0.046	0.034	0.044	0.223	−0.677
M/B_{t-1}	0.013	0.007	0.010	0.011	0.015	0.010	0.007	0.010	4.470	
Depreciation/TA_{t-1}	0.365	0.342	0.437	0.494	0.650	0.489	0.365	0.449	0.002	
$Ln(TA)_{t-1}$	0.049	0.014	0.010	0.010	0.012	0.013	0.006	0.016	3.937	
Fixed assets/TA_{t-1}	0.168	0.103	0.105	0.118	0.171	0.134	0.084	0.126	0.029	
R&D/TA_{t-1}	0.263	0.232	0.304	0.331	0.445	0.313	0.246	0.305	0.004	
Industry $median_{t-1}$	0.227	0.252	0.402	0.417	0.439	0.318	0.266	0.331	0.003	
Average RMSE	0.160	0.140	0.188	0.205	0.256	0.189	0.144			
Panel C. T=30										
EBIT/TA_{t-1}	0.022	0.021	0.031	0.035	0.056	0.027	0.021	0.030	0.223	−0.690
M/B_{t-1}	0.011	0.005	0.006	0.007	0.013	0.006	0.005	0.008	4.470	
Depreciation/TA_{t-1}	0.225	0.204	0.251	0.327	0.589	0.254	0.207	0.294	0.002	
$Ln(TA)_{t-1}$	0.050	0.008	0.006	0.005	0.009	0.011	0.003	0.013	3.937	
Fixed assets/TA_{t-1}	0.156	0.060	0.066	0.080	0.139	0.100	0.052	0.093	0.029	
R&D/TA_{t-1}	0.178	0.145	0.182	0.236	0.413	0.186	0.147	0.212	0.004	
Industry $median_{t-1}$	0.153	0.133	0.237	0.283	0.378	0.176	0.133	0.213	0.003	
Average RMSE	0.114	0.082	0.111	0.139	0.228	0.109	0.081			

panel lengths so that half of the firms have $T=6$ and half have $T=18$. We run 500 replications and present the results in Table 7. Reassuringly, the degree of imbalance has little impact on estimation accuracy for most methods. LSDVC continues to be the most accurate estimator while OLS is the least.

6.2. Missing observations

Actual Compustat datasets generally have missing data that require some observations for a given firm to be omitted from the analysis. We know of only one paper (Frank and Goyal (2009)) that examines the effect of such omissions, and their treatment is brief. We therefore investigate the issue here. For each of the 500 baseline balanced panels simulated with $T=12$, we randomly delete 10% of the firm-year observations.[13] We then estimate the coefficients with each of the seven methodologies across the modified data. Table 7 reports these results. Although the RMSEs are slightly increased for the advanced methodologies, randomly missing observations have little impact on most of the estimators.[14] LSDVC continues to be the best choice.

6.3. Dependent variable censoring or clustering

We next consider two possibilities that would generate clustered observations for the dependent variable. First, we evaluate the effects of a censored dependent variable, which is a particularly serious concern given the number of corporate finance fractional response variables such as leverage or payout ratios (Loudermilk (2007)). Iliev and Welch (2010) emphasize the importance of bounded dependent variables and Hovakimian and Li (2011) also raise the possibility of hard-wired mean reversion with censored data. To address this issue, we winsorize the top and bottom 4% of dependent variables for each of the 500 simulated datasets.[15] Results are reported in Table 7. Even with this level of censoring, LSDVC and BB remain the most accurate estimators. This conclusion is consistent with Papke and Wooldridge (2008),which shows linear models can provide reasonable estimates of average effects for fractional response dependent variables. This is an important finding for any empirical researchers using censored data — not just those specifically interested in capital structure adjustment models. While censoring somewhat exacerbates RMSE differences between $\gamma=0.2$ and 0.8 for some estimation methods, BB is unaffected and thus may be the best choice when the level of persistence is unknown.

Next, we construct data sets in which some firms' independent variables are not generated exclusively by the baseline model (5). We generate 500 datasets according to the baseline model (5), randomly select 10% of the firms and set the dependent variable to zero for all of those firms' observations. Such zero observations might reflect missing data, some type of selection issue, or an omitted variable.[16] The results are presented in the rightmost two columns of Table 7. This type of random clustering does not greatly affect the estimation methods' accuracies. AB, BB, and LSDVC perform reasonably well. LD4 is the most accurate when $\gamma=0.2$ but becomes unreliable with stronger persistence and thus should be avoided when γ could be large.

In summary, LSDVC and BB are the best choices across a range of data limitations. LSDVC is preferred with missing observations or unbalanced panels while BB has smaller RMSE when the dependent variable is censored or clustered at zero. Generally, both AB and LD are reasonable estimators, but LD4 is accurate only for dynamic panels with lower lag persistence.

7. Correlation in the errors

Based on the simulations thus far, LSDVC and Blundell Bond appear to be the most reliable estimators. However, each methodology has a known shortcoming. LSDVC assumes the regressors are exogenous (and independent of the error term) and BB instrumental variables are invalidated by the presence of second order serial correlation. We next introduce both issues into our simulations and compare the performance of all seven estimators.

7.1. Endogeneity

Endogeneity is a central issue for corporate empiricists (Roberts and Whited (2011)), and one of our best-performing estimation methods (LSDVC) assumes strictly exogenous regressors. To extend our analysis to include endogenous independent variables, we follow Wintoki et al., (2012) by generating three endogenous variables (out of the seven independent variables) as depending on prior within-firm realizations of the variable as well as both the lagged dependent variable (y_{t-1}) and the firm fixed effect (η). [17]

$$x_{ijt} = \rho_j x_{ijt-1} + \alpha_1 y_{it-1} + \alpha_2 \eta_i + \xi_{ijt}, \quad \text{for} \quad j = 1-3. \tag{6}$$

[13] Systematic missing data due to selection issues is beyond the scope of this analysis.
[14] Frank and Goyal (2009, page 23) similarly conclude that "it is remarkable how little change is observed" when they impute values for missing Compustat data.
[15] The choice of 4% (or 8% total) mimics the magnitude of censoring found in capital structure partial adjustment models where the dependent variable is limited to the (0, 1) range.
[16] For example, firms may elect zero payout or leverage for reasons other than those modeled and a two-stage model might be more appropriate.
[17] We thank an anonymous referee for encouraging us to investigate this issue.

Table 7
Panels with data limitations. For each of the base parameter variations ($T=12$, $\gamma=0.2$ or 0.8), 500 dynamic panels are generated using Compustat innovations. This table summarizes how each estimation method performs when used to estimate the model with each simulation. The "Lag RMSE" rows report the average RMSE of the estimated coefficients on the lagged dependent variable across the 500 simulations. The "Xs RMSE" rows report the average of the seven explanatory variables' estimated regressions, averaged across the 500 simulations. Numbers in **bold** (*italics*) are the **smallest** (*second-smallest*) RMSE for the lag or Xs, within each column.

		Panel imbalance		Missing data		Censored data		Cluster at zero	
	Lag value (γ):	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8
OLS	Lag RMSE	0.212	0.109	0.212	0.110	0.200	0.098	0.171	**0.026**
	Xs RMSE	0.149	0.152	*0.162*	0.164	0.150	0.158	0.166	0.269
Fixed effects	Lag RMSE	0.073	0.070	0.077	0.089	0.089	0.112	0.089	0.161
	Xs RMSE	**0.132**	**0.136**	**0.147**	**0.150**	**0.130**	**0.135**	**0.139**	*0.164*
Arellano Bond	Lag RMSE	0.017	0.028	0.025	0.072	0.022	0.036	*0.022*	*0.027*
	Xs RMSE	0.181	0.180	0.217	0.212	0.178	0.180	0.186	0.185
Blundell Bond	Lag RMSE	*0.015*	*0.016*	*0.022*	*0.025*	**0.018**	**0.015**	0.029	0.035
	Xs RMSE	0.202	0.202	0.239	0.240	0.195	0.199	0.201	0.200
Longest diff	Lag RMSE	0.034	0.036	0.036	0.051	0.030	0.045	0.028	0.037
	Xs RMSE	0.315	0.318	0.325	0.336	0.245	0.257	0.270	0.272
4 Period diff	Lag RMSE	0.032	0.124	0.036	0.127	0.040	0.140	**0.019**	0.114
	Xs RMSE	0.184	0.194	0.253	0.259	0.175	0.181	0.183	0.186
LSDVC	Lag RMSE	**0.013**	**0.009**	**0.017**	**0.020**	*0.019*	*0.030*	0.053	0.051
	Xs RMSE	*0.137*	*0.137*	0.164	*0.163*	*0.138*	*0.141*	*0.146*	**0.148**

The other four variables continue to be exogenous:

$$x_{ijt} = \rho_j x_{ijt-1} + \xi_{ijt}, \quad \text{for} \quad j = 4-7. \tag{6a}$$

The importance of the lag and the fixed effect in Eq. (6) are set to either $\alpha_1 = \alpha_2 = 0.01$ (low endogeneity) or $\alpha_1 = \alpha_2 = 0.05$ (high endogeneity).

To verify that this modification induces the desired endogeneity, we follow Wintoki et al. (2012) and use the Wooldridge (2009) test of exogeneity. The Wooldridge test adds future values ($t+1$) of independent variables to the baseline regression Eq. (5) and checks whether those values are statistically significant. Future values should not have any predictive power if the variables are exogenous. Indeed, in our high endogeneity data simulations, all three endogenous variables fail the Wooldridge test in each of the 500 simulations. In the low endogeneity data, at least one of the three endogenous variables fails the Wooldridge test in more than 80% of the simulated datasets. We also confirm that our baseline simulations (where the variables are not constructed to have endogeneity) only fail ~5% of the time, in line with the expected rejection of a true null.

Table 8 and the left column of Fig. 2 summarize the results with either high or low levels of endogeneity.[18] We report the average RMSE separately for the exogenous and endogenous variables. Table 8 shows that the endogenous variables' coefficients are generally estimated with greater error. While endogeneity has little effect on FE and BB, AB exhibits huge difficulties in estimating coefficients in the presence of even low endogeneity. Surprisingly, LSDVC still appears to be the best choice when there is low endogeneity and low persistence, despite the fact that it is not designed to work with endogenous regressors. FE is accurate in estimating both the exogenous and endogenous Xs, but not the lagged dependent variable. Unfortunately, Table 9 shows that FE estimate accuracy deteriorates when the panels are unbalanced. BB remains the best option for higher levels of endogeneity if the lagged dependent variable is of interest.

Next we extend our analysis of endogenous regressors by reintroducing the common corporate finance data limitations described in Section 6. We simulate panels with high endogeneity ($\alpha_1 = \alpha_2 = 0.05$) in three of the independent variables and then evaluate the impact of panel imbalance, missing data, censored data, and dependent variable clustering at zero. Table 9 presents these results.

The first two columns of Table 9 show how the coefficient estimates are affected when panel length is unbalanced. As in Section 6.1, we induce a high degree of imbalance. The average panel length remains 12, but half of the observations are $T=6$ and half are $T=18$. Comparing these results to the last two columns of Table 8 (the high endogeneity results), we quickly see that FE is no longer the dominant means to estimate either exogenous or endogenous X coefficients. Instead we see that BB and LD appear to be the most robust methodologies for unbalanced panels with endogenous variables. LSDVC continues to be a reasonable choice as well. However, the endogenous variables' mean RMSE is so large that it would be difficult to make reliable inferences from these estimates. Looking closer, two things are clear. The RMSE are almost uniformly higher (and often much higher) for an unbalanced panel. This implies that there isn't a monotonic relationship between panel length and estimate accuracy. The shorter panels are very difficult to estimate, regardless of methodology. Further, a highly persistence lag structure ($\gamma=0.8$) is deleterious to almost every approach and makes it virtually impossible to get accurate estimates for endogenous variables. These columns indicate that estimating a highly unbalanced panel with endogeneity should be undertaken with extreme caution.

[18] Note that the y-axis scaling of Fig. 2 differs from Fig. 1 to accommodate the higher observed RMSE.

Table 8
Panels with endogenous explanatory variables. For both low and higher levels of endogeneity and with both adjustment processes ($\gamma = 0.2$ or 0.8), 500 dynamic panels are generated using Compustat innovations for a $T = 12$ balanced panel. This table summarizes how each estimation method performs when used to estimate the model with each simulation. The "Lag RMSE" rows report the average RMSE of the estimated coefficients on the lagged dependent variable across the 500 simulations. The "Exo Xs RMSE" rows report the average of the four exogenous explanatory variables' estimated regressions, averaged across the 500 simulations. The "Endo Xs RMSE" rows report the average of the three endogenous explanatory variables' estimated regressions, averaged across the 500 simulations. Numbers in **bold** (*italics*) are the **smallest** (*second-smallest*) RMSE for the lag, exogenous Xs, or endogenous Xs, within each column.

		Lower endogeneity		Higher endogeneity	
	Lag value (γ):	0.2	0.8	0.2	0.8
OLS	Lag RMSE	0.190	0.093	0.156	0.145
	Exo Xs RMSE	0.177	0.182	0.184	0.180
	Endo Xs RMSE	0.620	0.399	1.901	1.317
Fixed effects	Lag RMSE	0.077	0.087	0.069	0.083
	Exo Xs RMSE	*0.148*	*0.154*	**0.149**	**0.153**
	Endo Xs RMSE	**0.127**	**0.127**	**0.149**	**0.129**
Arellano Bond	Lag RMSE	0.036	0.038	0.758	0.649
	Exo Xs RMSE	0.222	0.219	0.400	0.360
	Endo Xs RMSE	1.166	1.145	5.292	4.666
Blundell Bond	Lag RMSE	*0.016*	*0.021*	**0.031**	**0.029**
	Exo Xs RMSE	0.226	0.226	0.226	0.227
	Endo Xs RMSE	0.183	*0.186*	*0.211*	*0.224*
Longest diff	Lag RMSE	0.030	0.043	*0.066*	0.220
	Exo Xs RMSE	0.279	0.282	0.280	0.289
	Endo Xs RMSE	0.252	0.392	0.341	1.303
4 Period diff	Lag RMSE	0.035	0.138	0.091	0.335
	Exo Xs RMSE	0.185	0.198	0.187	0.207
	Endo Xs RMSE	0.191	0.371	0.407	1.339
LSDVC	Lag RMSE	**0.014**	**0.014**	0.097	*0.077*
	Exo Xs RMSE	**0.147**	**0.150**	*0.156*	*0.164*
	Endo Xs RMSE	*0.169*	0.459	0.662	0.365

The next two columns in Table 9 focus on randomly missing data in a panel with some endogenous regressors. BB is the best choice regardless of the persistence. Having returned to a balanced panel, we again see that FE is a reasonable estimator for the independent coefficients when there is low persistence ($\gamma = 0.2$). However, it is far less accurate when persistence is high and some observations are missing. LSDVC is accurate for the exogenous regressors but less accurate for the lagged dependent variable in the presence of endogeneity.

The last four columns in Table 9 illustrate how dependent variable limitations affect the methodologies in the presence of endogenous variables. For a censored dependent variable, BB is the most accurate estimator overall, with FE again being a good choice if the coefficients of interest are the independent variables and not the lag. However, FE outperforms BB with fixed panel lengths when the dependent variable is clustered. These results imply a tradeoff between the two methodologies for empiricists interested in capital structure, payout, or other areas with clustering at zero and a dynamic panel. Clustering and panel length distributions should both be considered in choosing between FE and BB estimators.

7.2. Serially correlated regression residuals

Discussions of the AB and BB GMM-style estimators emphasize that second order serial correlation in the residuals renders their instruments invalid.[19] In order to determine the empirical magnitude of this potential problem, we modify the error term in Eq. (7) and generate data with $\delta_1 = 0.10$ and $\delta_2 = 0.05$ (these values are in line with the level of persistence found in capital structure data). Comparing the results in Table 10 and the right column of Fig. 2 illustrates the detrimental effect of second order serial correlation on the GMM estimators. Consistent with theoretical predictions, the GMM estimators perform worse when the regression residuals are serially correlated (comparing the left columns of Table 10 to the middle columns of Table 5). Interestingly, LD – designed to sidestep the second order serial correlation concern of AB and BB – also performs worse at least in balanced panels with $T = 12$. It is worthwhile to note that this form of correlation in the errors has little impact on the estimation of the exogenous regressors' coefficients. Lastly, there is a sharp difference between the RMSEs associated with a true $\gamma = 0.8$ and those associated with $\gamma = 0.2$: a smaller γ (lower persistence in the dependent variable) generates higher estimation errors. This is reassuring for the capital structure literature, for which adjustment speeds $(1 - \gamma)$ appear to lie in the neighborhood of 15–25% range (Lemmon et al. (2008), Huang and Ritter (2009)). GMM estimators thus seem well-suited to estimating capital structure models, even in the presence of second-order serial correlation.

All else equal, FE and LSDVC are the most accurate estimators in the presence of second order serial correlation if no endogeneity exists. However as endogeneity increases, FE dominates the alternative choices given our simulation parameters. The

[19] Wooldridge (2009) presents a test for serial correlation in the errors.

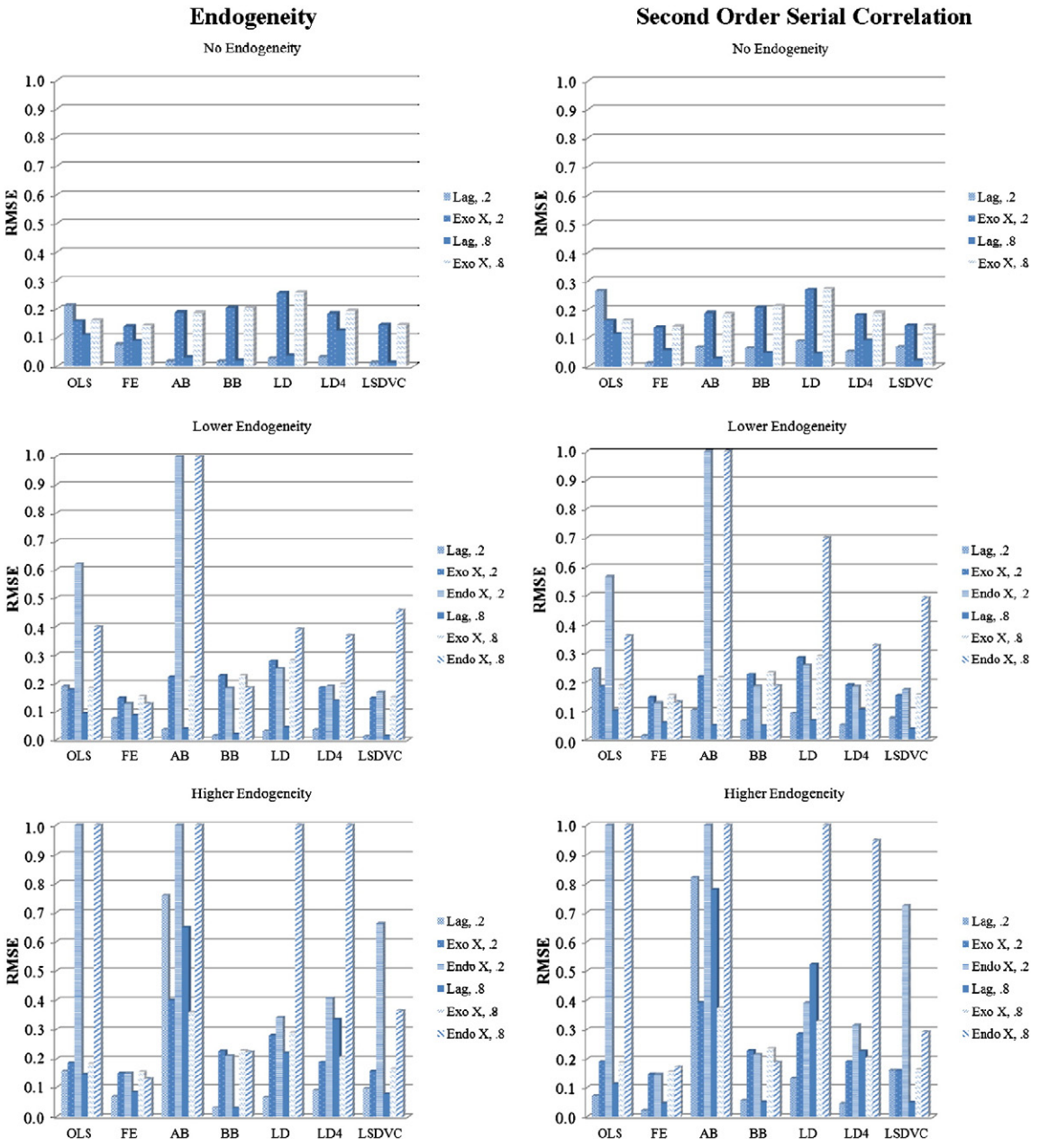

Fig. 2. Balanced panels with endogenous variables and second order serial correlation. Source: Tables 5, 8, and 10. All data uses $T=12$ with Compustat innovations. For high and low persistence ($\gamma=0.2$ or 0.8), we present the lag dependent variable RMSE (*Lag*), the average RMSE for the four exogenous variables (*Exo X*), and the average RMSE for the three endogenous variables (*Endo X*). The left column summarizes the simulations across varying levels of endogeneity with the first graph taken from Fig. 1 ($T=12$, Compustat). The right column presents the impact of second order serial correlation as the level of endogeneity varies. The y axis is censored for the purpose of readability.

performance of FE warrants some discussion. The five advanced estimators were created to address the known downward bias of FE estimates in dynamic panels. However, the negative bias depends on the correlation of the transformed lag and the transformed error (Nickell, 1981). As modeled in our simulations, the second order serial correlation changes this relationship between the two transformed variables and mitigates the bias. Unfortunately, we cannot conclude that FE as always the best estimator in the presence of second order serial correlation, because our simulations have not explored the impact of alternative levels of persistence in the regression residuals. Moreover, serial correlation can lead to incorrect standard errors in FE (Wooldridge (2009)).

7.3. Implications

We conclude that endogeneity and serial correlation pose more than just theoretical challenges. AB and LSDVC are particularly affected by endogeneity while all of the GMM-style estimators are affected to some extent by second order serial correlation. FE

Table 9
Panels with endogenous variables and data limitations. For each of the base parameter variations ($T=12$, $\gamma=0.2$ or 0.8), 500 dynamic panels are generated using Compustat innovations and high endogeneity (or $\alpha_1=\alpha_2=0.05$). This table summarizes how each estimation method performs when used to estimate the model with each simulation. . The "Lag RMSE" rows report the average RMSE of the estimated coefficients on the lagged dependent variable across the 500 simulations. The "Exo Xs RMSE" rows report the average of the four exogenous explanatory variables' estimated regressions, averaged across the 500 simulations. The "Endo Xs RMSE" rows report the average of the three endogenous explanatory variables' estimated regressions, averaged across the 500 simulations. Numbers in **bold** (*italics*) are the **smallest** (*second-smallest*) RMSE for the lag, exogenous Xs, or endogenous Xs, within each column.

		Panel imbalance		Missing data		Censored data		Cluster at zero	
	Lag value (γ):	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8
OLS	Lag RMSE	0.167	*0.644*	0.154	0.285	0.134	**0.021**	0.084	0.129
	Exo Xs RMSE	0.204	**0.261**	0.192	0.201	0.182	0.182	0.186	0.186
	Endo Xs RMSE	1.948	3.989	1.898	2.070	1.658	0.439	0.876	**0.180**
Fixed effects	Lag RMSE	0.203	0.688	0.142	0.276	0.076	0.084	**0.036**	*0.061*
	Exo Xs RMSE	**0.182**	0.266	**0.166**	*0.188*	**0.137**	**0.142**	**0.149**	**0.152**
	Endo Xs RMSE	1.140	3.971	*0.276*	1.005	**0.130**	**0.127**	0.299	*0.229*
Arellano Bond	Lag RMSE	0.200	0.706	0.694	0.396	0.679	0.439	0.639	0.622
	Exo Xs RMSE	0.333	0.331	0.409	0.331	0.346	0.293	0.344	0.334
	Endo Xs RMSE	1.651	*3.788*	5.094	3.732	4.515	3.457	4.346	4.199
Blundell Bond	Lag RMSE	**0.114**	0.705	**0.041**	**0.037**	**0.032**	*0.045*	0.050	0.092
	Exo Xs RMSE	0.241	0.334	0.267	0.268	0.215	0.218	0.220	0.222
	Endo Xs RMSE	**0.988**	4.168	**0.259**	**0.272**	*0.201*	*0.168*	**0.206**	0.343
Longest diff	Lag RMSE	0.129	**0.509**	*0.082*	0.264	*0.062*	0.135	0.070	0.257
	Exo Xs RMSE	0.412	0.457	0.356	0.373	0.266	0.285	0.294	0.305
	Endo Xs RMSE	1.094	**3.131**	0.415	1.519	0.301	0.867	0.360	1.505
4 Period diff	Lag RMSE	0.248	0.886	0.101	0.345	0.078	0.167	*0.039*	0.157
	Exo Xs RMSE	0.234	0.348	0.251	0.265	0.176	0.180	0.185	0.188
	Endo Xs RMSE	1.486	4.878	0.461	1.415	0.319	0.387	*0.223*	0.406
LSDVC	Lag RMSE	*0.125*	0.655	0.094	*0.054*	0.082	0.049	0.126	**0.046**
	Exo Xs RMSE	*0.185*	*0.263*	*0.174*	**0.179**	*0.145*	*0.152*	*0.154*	*0.157*
	Endo Xs RMSE	*1.041*	3.839	0.661	*0.670*	0.475	0.613	0.625	0.492

often provides reliable independent variable estimates and is superior overall for dependent variable clustering, but is highly sensitive to panel length and imbalance. BB appears to be the best choice in the presence of endogeneity and even (surprisingly) second order serial correlation if the dataset includes shorter panels.

8. Conclusion: empirical implications for researchers

This paper examines the performance of seven econometric methodologies for estimating dynamic panel models in "realistic" corporate finance datasets. We consider two methods whose results are thought to be generally biased (OLS and fixed effects (FE)), plus five "advanced" estimation methods (Arellano Bond (AB), Blundell Bond (BB), longest differencing (LD), four period differencing (LD4), and least squares dummy variable correction (LSDVC)). Although each of the advanced methodologies has been tested using simple Monte Carlo studies, this is the first paper to examine their accuracy under the sort of statistical conditions contained in large corporate finance databases such as unbalanced panels and endogenous regressors. We emphasize that the long difference estimators have been formally derived only for balanced panels, and that our implementations here (LD, LD4) should be considered preliminary.

We find that the advanced estimators generally work as advertised in estimating the coefficient on a dynamic panel's lagged dependent variable when the independent variables are exogenous. Most of the estimators provide small root-mean squared errors (RMSEs) when estimating a dependent variable's persistence, regardless of the true value (high or low). As expected, OLS and FE do poorly, on average, in these estimations and FE is affected particularly by shorter panels. More noteworthy is the poor performance of LD4 when the dependent variable is highly persistent ($\gamma=0.8$). LSDVC, AB, and BB have the lowest RMSE in estimating the lagged dependent variable and are reasonably accurate for the independent variables. FE often is the most accurate with respect to the exogenous variables, but exhibits much higher errors for the lag.

While the preliminary analysis shows that LSDVC is the most accurate estimator across a range of data limitations, theory predicts that endogenous explanatory variables will reduce the quality of the LSDVC estimates. We find this to be true, but the damage is limited to the estimated coefficients on the endogenous variables: LSDVC remains accurate for the lagged dependent variable and exogenous regressors. FE exhibits low RMSE for the endogenous variables, but is not very accurate for the lagged dependent variable. BB is reliable regardless of the level of endogeneity or dependent variable persistence and should be the default choice under these conditions, particularly if the lag coefficient is of interest.

We confirm theoretical predictions that AB and BB perform worse when the true model's residuals exhibit second-order serial correlation but the change is not dramatic. The increased estimation error is most apparent for the lagged dependent variable's coefficient; the explanatory variables' RMSEs are largely unaffected. Somewhat surprisingly, FE provides the most accurate estimates in the presence of second-order serial correlation, with BB a close second. The BB estimates sometimes outperform FE (estimating a

Table 10
Panels with second order serial correlation. For varying levels of endogeneity and with both adjustment processes ($\gamma = 0.2$ or 0.8), 500 dynamic panels are generated using Compustat innovations and second order serial correlation for a $T = 12$ balanced panel. Endogeneity is included for the first three of the seven independent variables. This table summarizes how each estimation method performs when used to estimate the model with each simulation. The "Lag RMSE" rows report the average RMSE of the estimated coefficients on the lagged dependent variable across the 500 simulations. The "Exo Xs RMSE" rows report the average of the four exogenous explanatory variables' estimated regressions, averaged across the 500 simulations. The "Endo Xs RMSE" rows report the average of the three endogenous explanatory variables' estimated regressions, averaged across the 500 simulations. Numbers in **bold** (*italics*) are the **smallest** (*second-smallest*) RMSE for the lag, exogenous Xs, or endogenous Xs, within each column.

		All exogenous X		Lower endogeneity		Higher endogeneity	
	Lag value (γ):	0.2	0.8	0.2	0.8	0.2	0.8
OLS	Lag RMSE	0.265	0.115	0.244	0.100	0.071	0.112
	Exo Xs RMSE	0.161	0.162	0.183	0.187	0.190	0.187
	Endo Xs RMSE			0.565	0.361	1.711	1.159
Fixed effects	Lag RMSE	**0.013**	0.058	**0.014**	0.058	**0.021**	**0.047**
	Exo Xs RMSE	**0.137**	**0.141**	**0.146**	**0.153**	**0.146**	**0.153**
	Endo Xs RMSE			**0.127**	**0.129**	**0.146**	**0.170**
Arellano Bond	Lag RMSE	0.069	0.028	0.102	0.049	0.820	0.777
	Exo Xs RMSE	0.189	0.186	0.218	0.216	0.392	0.374
	Endo Xs RMSE			1.135	1.113	5.194	5.023
Blundell Bond	Lag RMSE	0.066	*0.048*	0.065	*0.048*	0.056	0.051
	Exo Xs RMSE	0.207	0.213	0.227	0.234	0.228	0.235
	Endo Xs RMSE			0.185	*0.187*	*0.215*	*0.188*
Longest diff	Lag RMSE	0.089	0.046	0.090	0.066	0.131	0.524
	Exo Xs RMSE	0.270	0.273	0.285	0.290	0.285	0.328
	Endo Xs RMSE			0.258	0.701	0.391	3.008
4 Period diff	Lag RMSE	*0.054*	0.092	*0.051*	0.104	*0.046*	0.226
	Exo Xs RMSE	0.182	0.189	0.189	0.199	0.190	0.205
	Endo Xs RMSE			0.184	0.327	0.316	0.948
LSDVC	Lag RMSE	0.070	**0.022**	0.074	**0.036**	0.158	*0.050*
	Exo Xs RMSE	*0.144*	*0.144*	*0.152*	*0.155*	*0.158*	*0.162*
	Endo Xs RMSE			*0.174*	0.493	0.723	0.292

highly persistent lag coefficient), are unaffected by panel imbalance, and are consistent across a range of endogeneity in the presence of serial correlation.

While LSDVC is the most accurate estimator in the absence of endogenous independent variables and second order serial correlation, BB and FE are the most accurate estimators when those complications exist. BB is more affected by dependent variable clustering at zero and second order serial correlation and FE performs poorly in short (or unbalanced) panels and is sensitive to highly persistent data.

Across the "Compustat" type explanatory variables, estimation errors are inversely related to the variances of the variables' innovations, which vary substantially. The estimated coefficients on explanatory variables with the largest innovation variances have RMSEs similar to those on independent, normally-distributed independent variables. However, the RMSEs for lower-variance Compustat variables can be an order of magnitude larger. Although these RMSEs decline as panel length increases, the estimation errors remain larger even for panels that are 30-years long. This finding suggests that researchers should examine an explanatory variables' within-firm variance before assessing its economic effect on the dependent variable in a dynamic panel model.

Clearly, we have not evaluated dynamic panel model estimation techniques under all conceivable conditions. However, the simulation results reported here should assist corporate finance researchers in estimating such models. Many of the problematic dataset properties can be identified through tests (endogeneity, serial correlation) or by inspection (variable value clustering, short panels). Combined with a dataset's known characteristics, the results of this paper can guide empiricists working with dynamic panels.

Appendix A. Implementation of long differencing in unbalanced panels

Hahn, et al. (2007) introduces a new instrumental variable estimator for dynamic panels based on long differencing work by Griliches and Hausman (1986). Their estimation method can be summarized as follows. Just as first differencing can yield valid instruments for the lagged dependent variable (Arellano and Bond (1991)), so too can longer differencing windows. The goal of the Hahn, et al. paper was to circumvent limits that the second order serial correlation imposes on AB and BB. Consider a panel with "t" observations on the same firm. Instead of first-differencing, take the longest difference available:

$$y_{it} - y_{i2} = \lambda(y_{it-1} - y_{i1}) + \alpha(x_{it} - x_{i2}) + (\varepsilon_{it} - \varepsilon_{i2}). \tag{9}$$

Hahn, et al. observes that y_{i1} is correlated with the lagged difference of the dependent variable and, thus, a valid instrument for the differenced lag. Further, they appeal to the work of Hausman and Taylor (1983) and Ahn and Schmidt (1995) to state that residuals from this model also would serve as valid instruments. The long difference estimator they propose, therefore, is iterative. Initially, only y_{i1} is used as instrument for the model estimation. Then the residuals are estimated and the model is

re-estimated, this time using both y_{i1} and the estimated residuals are instruments. Now, the residuals can be calculated again and the model is estimated for a third time using y_{i1} and the new residuals. The coefficients of the third iteration are reported.

Since Hahn et al. treat only balanced panels, there is no tested application for unbalanced data. To implement a LD estimator with unbalanced panels, we utilize the longest available difference for each firm (this means that the differencing interval varies across firms). When Huang and Ritter (2009) confronted the problem of unbalanced panels, they chose to impose an equal differencing interval on all firms.

$$y_{it} - y_{it-k} = \lambda(y_{it-1} - y_{it-k-1}) + \alpha(x_{it} - x_{it-k}) + (\varepsilon_{it} - \varepsilon_{it-k}). \tag{10}$$

While their paper includes 4, 8, and higher level differences, we only include the four period differencing as that excludes the least data from the estimation. Our LD implementation for unbalanced panels includes all firms by defining each "long difference" as the length of the firm's panel. LD4 drops firms that have fewer than five observations. Following Hahn, Hausman, and Kuersteiner, we use y_{i0} as an instrument for the lagged difference for LD. For LD4, we use y_{it-5} as the initial instrument.

Appendix B. Overview of simulation parameter choices

Our method for generating simulated datasets permits the introduction of various features to the data for which we seek to estimate the main specification Eq. (5). Specifically, appropriate parameter values in Eqs. (6) and (7) can introduce serially correlated regression residuals, correlated explanatory variables, and endogeneity in some of the explanatory variables.

The general model for data generation is:

$$y_{it} = \gamma y_{it-1} + \sum_{j=1}^{7} \sum \beta_j x_{jit} + \eta_i + \varepsilon_{it} \tag{5}$$

$$x_{ijt} = \rho_j x_{ijt-1} + \alpha_1 y_{it-1} + \alpha_2 \eta_i + \xi_{ijt}, \quad j = 1, 7 \tag{6}$$

$$\varepsilon_{it} = \delta_1 \varepsilon_{i,t-1} + \delta_2 \varepsilon_{i,t-2} + \omega_{it}. \tag{7}$$

The main parameters of interest in estimating dynamic panel models are γ and the β_j. We set $\gamma = 0.2$ or 0.8 and all seven $\beta_i = 0.2$ (Arellano and Bond (1991), Kiviet (1995), and Judson and Owen (1999) select similar values). Firm fixed effects (η_i) are uniformly distributed over the [−1, 1], which avoids fixed effects clustered near zero. Until Section 7, we assume exogenous regressors ($\alpha_1 = \alpha_2 = 0$ in Eq. (6)) and serially independent residuals ($\delta_1 = \delta_2 = 0$ in Eq. (7)).

B.1. iid Innovations

Each of the seven independent variables is constructed with a persistence (ρ_j) corresponding to the within-firm persistence of a corporate variable, estimated for the CRSP/Compustat universe (excluding financials and utilities) over the period 1962–2004. These variables are the ones used in our specification for a dynamic leverage model: EBIT/TA, M/B, Depreciation/TA, Ln(TA), Fixed Assets/TA, R&D/TA, or Industry Median MDR. The resulting persistence values are reported in Table 3A.

$$\rho_1 = .455; \rho_2 = .259; \rho_3 = .085; \rho_4 = .844; \rho_5 = .254; \rho_6 = .197; \rho_7 = .690.$$

The iid data sets are created with independent, normally distributed residual terms.

$$\varepsilon_{it} \sim N(0,1) \qquad \xi_{it} \sim N(0,1) \qquad \omega_{it} \sim N(0,1)$$

B.2. Compustat-style innovations

The independent variable innovations (ξ_{it}) are drawn from a multivariate normal distribution using a covariance matrix of seven corporate variables computed from the set of CRSP-Compustat firms during 1962–2004 (excluding financial and utility firms). The matrix is reported in Table 3B. Unlike the iid case, Compustat-style explanatory variables are correlated with one another and differ in their innovation variances (ξ_{it}).

B.3. Endogeneity

Using the Compustat-style process for generating innovations, we specify some endogeneity for three of the independent variables while leaving the others as exogenous:

$$x_{ijt} = \rho_j x_{ijt-1} + \alpha_1 y_{it-1} + \alpha_2 \eta_i + \xi_{ijt}, \quad \text{for} \quad j = 1, 2, 3 \tag{6}$$

$$x_{ijt} = \rho_j x_{ijt-1} + \xi_{ijt}, \quad \text{for} \quad j = 4, 5, 6, 7. \tag{6a}$$

Datasets are constructed with three alternative levels of endogeneity: "Low" ($\alpha_1 = \alpha_2 = 0.01$), "High" ($\alpha_1 = \alpha_2 = 0.05$), and "None" ($\alpha_1 = \alpha_2 = 0.00$).

Appendix 3. Second order serial correlation

Second order serial correlation is introduced to the error term of the Eq. (5) by setting

$$\delta_1 = 0.10; \delta_2 = 0.05$$

in Eq. (7). The rest of the data is based on the Compustat-style, both with and without the presence of endogenous variables.

References

Ahn, S.C., Schmidt, P., 1995. Efficient estimation of models for dynamic panel data. J. Econometrics 68, 5–27.
Andres, C., Betzer, A., Goergen, M., Renneboog, L., 2009. Dividend policy of German firms: a panel data analysis of partial adjustment models. J. Empir. Financ. 16, 175–187.
Arellano, M., Bond, S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. Rev. Econ. Stud. 58, 277–297.
Arellano, M., Bover, O., 1995. Another look at the instrumental variable estimation of error-components models. J. Econometrics 68, 29–51.
Baltagi, B.H., 2008. Econometric Analysis of Panel Data. John Wiley and Sons, West Sussex.
Blundell, R., Bond, S., 1998. Initial conditions and moment restrictions in dynamic panel data models. J. Econometrics 87, 115–143.
Bond, S., 2002. Dynamic panel data models: a guide to micro data methods and practice, practice. Port. Econ. J. 1, 141–162.
Bond, S., Meghir, C., 1994. Dynamic investment models and the firm's financial policy. Rev. Econ. Stud. 61, 197–222.
Bond, S., Elston, J.A., Mairesse, J., Mulkay, B., 2003. Financial factors and investment in Belgium, France, Germany, and the United Kingdom: a comparison using company panel data. Rev. Econ. Stat. 85, 153–165.
Bound, J., Jaeger, D.A., Baker, R.M., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. J. Am. Stat. Assoc. 90, 443–450.
Bruno, G.S.F., 2005. Approximating the bias of the LSDV estimator for dynamic unbalanced panel data model. Econ. Lett. 87, 361–366.
Chang, X., Dasgupta, S., 2009. Target behavior and financing: how conclusive is the evidence? J. Financ. 64, 1767–1796.
Dittmar, A.K., Duchin, R., 2010. The Dynamics of Cash. University of Michigan Working Paper.
Elsas, R., Florysiak, D., 2011. Dynamic Capital Structure Adjustment and the Impact of Fractional Dependent Variables. Available at SSRN: http://ssrn.com/abstract=1632362.
Fama, E.F., French, K.R., 1997. Industry costs of equity. J. Financ. Econ. 43, 153–193.
Fama, E.F., French, K.R., 2002. Testing trade-off and pecking order theories of capital structure. Rev. Financ. Stud. 15, 1–33.
Faulkender, M.W., Flannery, M.J., Hankins, K.W., Smith, J., 2012. Cash flows and leverage adjustments. J. Financ. Econ. 103, 632–646.
Fischer, E.O., Heinkel, R., Zechner, J., 1989. Dynamic capital structure choice: theory and tests. J. Financ. 44, 19–40.
Flannery, M.J., Rangan, K.P., 2006. Partial adjustment toward target capital. J. Financ. Econ. 41, 41–73.
Frank, M.Z., Goyal, V.K., 2009. Capital structure decisions: which factors are reliably important? Financ. Manage. 38, 1–37.
Gormley, T.A., Matsa, D.A., 2012. Common Errors: How to (and Not to) Control for Unobserved Heterogeneity. University of Pennsylvania Working Paper.
Griliches, Z., Hausman, J., 1986. Errors in variables in panel data. J. Econometrics 31, 93–118.
Hahn, J., Hausman, J., Kuersteiner, G., 2007. Long difference instrumental variables estimation for dynamic panel models with fixed effects. J. Econometrics 140, 574–617.
Hausman, J.A., Taylor, W.E., 1983. Identification in linear simultaneous equations models with covariance restrictions: an instrumental variables interpretation. Econometrica 51, 1527–1550.
Hovakimian, A., Li, G., 2011. In search of conclusive evidence: how to test for adjustment to target capital structure. J. Corp. Financ. 17, 33–44.
Huang, R., Ritter, J., 2009. Testing theories of capital structure and estimating speed of adjustment. J. Financ. Quant. Anal. 44, 237–271.
Iliev, P., Welch, I., 2010. Reconciling Estimates of the Speed of Adjustment of Leverage Ratios. Working Paper.
Judson, R.A., Owen, A.L., 1999. Estimating dynamic panel data models: a guide for macroeconomists. Econ. Lett. 65, 9–15.
Kiviet, J.F., 1995. On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. J. Econometrics 68, 53–78.
Lemmon, M.L., Roberts, M.R., Zender, J.F., 2008. Back to the beginning: persistence and the cross-section of corporate capital structure. J. Financ. 63, 1575–1608.
Loudermilk, M.S., 2007. Estimation of fractional dependent variables in dynamic panel data models with an application to firm dividend policy. J. Bus. Econ. Stat. 25, 462–472.
Machin, S., Reenen, J.V., 1993. Profit margins and the business cycle: evidence from UK manufacturing firms. J. Ind. Econ. 41 (1), 29–50.
MacKay, P., Phillips, G.M., 2005. How does industry affect firm financial structure? Rev. Financ. Stud. 18, 1433–1466.
Nelson, C.R., Startz, R., 1990. The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. J. Bus. 63, 125–140.
Nerlove, M., 1967. Experimental evidence on the estimation of dynamic economic relations from a time series of cross sections. Econ. Stud. Quart. 18, 42–74.
Nickell, S., 1981. Biases in dynamic models with fixed effects. Econometrica 49, 1417–1426.
Ozkan, A., 2000. An empirical analysis of corporate debt maturity structure. Eur. Financ. Manage. 6 (2), 197–212.
Papke, L.E., Wooldridge, J.M., 2008. Panel data methods for fractional response variables with an application to test pass rates. J. Econometrics 145, 121–133.
Petersen, M.A., 2009. Estimating standard errors in finance panel data sets: comparing approaches. Rev. Financ. Stud. 22, 435–480.
Roberts, M.R., Whited, T.M., 2011. Endogeneity in empirical corporate finance. In: Constantinides, George, Harris, Milton, Stulz, Rene (Eds.), Handbook of the Economics of Finance Volume 2. Elsevier.
Welch, I., 2004. Capital structure and stock returns. J. Polit. Econ. 11 (2), 106–131.
Wintoki, M.B., Linck, J.S., Netter, J., 2012. Endogeneity and the dynamics of internal corporate governance. J. Finan. Econ. 105, 581–606.
Wooldridge, J.M., 2009. An Introduction to Econometrics. South-Western Cengage Learning.