Received: 22 January 2023 | Revised: 1 May 2023 | Accepted: 4 June 2023

DOI: 10.1002/jcaf.22647

RESEARCH ARTICLE

The Journal of Corporate Accounting & Finance WILEY

The selection of control variables in capital structure research with machine learning

Control variables in capital structure

Rumeysa Bilgin

Department of Business Administration Management, Entrepreneurship and Leadership Research and Application Center, Istanbul Sabahattin Zaim University, Istanbul, Turkey

Correspondence
Rumeysa Bilgin, Department of Business Administration Management, Entrepreneurship and Leadership Research and Application Center, Istanbul Sabahattin Zaim University, Istanbul, Turkey.
Email: rumeysa.bilgin@izu.edu.tr

Abstract

The previous literature on capital structure has produced plenty of potential determinants of leverage over the last decades. However, their research models usually cover only a restricted number of explanatory variables, and many suffer from omitted variable bias. This study contributes to the literature by advocating a sound approach to selecting the control variables for empirical capital structure studies. We applied linear LASSO inference approaches to evaluate the marginal contributions of three proposed determinants; cash holdings, non-debt tax shield, and current ratio. While some studies did not use these variables in their models, others obtained contradictory results. Our findings have revealed that cash holdings, current ratio, and non-debt tax shield are crucial factors that substantially affect the leverage decisions of firms and should be controlled in empirical capital structure studies.

KEYWORDS

determinants of capital structure, LASSO inference, leverage ratio

1 | INTRODUCTION

Factors affecting firms' capital structure decisions have attracted the attention of corporate finance researchers over the last half-century. Early findings reveal that the leverage ratio decreases with firm profitability and growth potential and increases with firm size and tangibility (Rajan & Zingales, 1995). More recently, researchers focused on testing new variables' role in capital structure decisions. These efforts have revealed other potential firm, industry- and country-specific determinants of leverage, such as firm age, industry munificence, inflation, financial orientation, and bank concentration (Antoniou et al., 2008; Baum et al., 2017; Bilgin & Dinc, 2019; De Jong et al., 2008; Frank & Goyal, 2009; Gonzalez & Gonzalez, 2008; Kayo & Kimura, 2011; Kieschnick & Moussawi, 2018; Lim et al., 2020). However, the empirical studies are far from providing comparable results since there is no consensus on the correct model specification, the estimation of the leverage ratio, and the selection of control variables. Although the previous literature discusses the first two issues to some extent, it shows little interest in selecting control variables. This study contributes to the literature by applying a sound approach to this selection process.

When the number of explanatory variables is close to the number of observations, conventional econometric models, such as OLS, overfit the noise (Nagel, 2021). As a result, capital structure models usually cover only a restricted number of explanatory variables and have limited predictive power. Besides, the control variables of these models are generally selected arbitrarily depending on data availability and p-hacking. Model misspecification problems (i.e., omitting relevant variables or including irrelevant ones) may cause biased and inefficient estimates and

incorrect *p*-values when the control variables are chosen based on the significance of their coefficients after trying different combinations. The issue of p-hacking is not specific to the finance literature. Instead, it is prevalent in almost all academic disciplines that rely on statistical inference (Head et al., 2015). Hypothesis testing is employed to detect the significance of the model variables in these disciplines. The researchers determine the significance level of these tests subjectively, and various combinations of variables are tried to create a set with significant *p*-values (Simmons et al., 2011).

Furthermore, capital structure research must enlarge its toolkit with new methods to overcome drawbacks and validate previous studies' findings. Even though theory-driven econometric methods are indispensable in empirical studies, researchers can discover new perspectives with more data-driven techniques. Machine Learning (ML) can be a good option. It is the state-of-the-art approach for analyzing high-dimensional data sets in finance (Gogas & Papadimitriou, 2021; Rundo et al., 2019). The data-driven ML methods can solve the problem of selecting the best control variables among a large set of possible ones and improve the model specification process.

This study contributes to the literature by advocating a sound approach to deciding the control variables for empirical capital structure studies. We employed the double selection LASSO (Belloni et al., 2012), partialing-out LASSO (Belloni et al., 2014; Chernozhukov et al., 2015a, 2015b) and cross-fit partialling-out LASSO of Chernozhukov et al. (2018) to evaluate the marginal contributions of cash holdings, non-debt tax shield, and current ratio as determinants of capital structure. These variables are selected because of the previous studies' inconclusive results on their role as capital structure determinants. Some researchers used them in their model specifications and obtained contradictory results, while others did not consider them as determinants of leverage (i.e., Chakrabarti & Chakrabarti, 2019;Czerwonka & Jaworski, 2022; Loncan & Caldeira, 2014; Panda et al., 2021; Pathak et al., 2021; Poornima & Kumar, 2022; Zhu et al., 2021). Our findings have revealed that cash holdings, current ratio, and non-debt tax shield are crucial factors that substantially affect the leverage decisions of firms and should be controlled in empirical capital structure studies. To the authors' knowledge, the current study represents the first application of the LASSO inference methods in capital structure research.

The rest of the paper is organized as follows. The second section reviews the relevant literature; the third introduces the data set and research methodology; and the fourth presents the findings. Robustness tests are described in section five. Lastly, section six concludes.

2 | LITERATURE REVIEW

2.1 | Capital structure

The empirical literature on capital structure shows extensive diversity in the model specifications and research methodologies. Some researchers employ linear regression models to analyze cross-sectional data structures (e.g., De Jong et al., 2008; Rajan & Zingales, 1995). Others prefer OLS-based static panel estimators (e.g., Alves & Ferreira, 2011) or dynamic approaches like system GMM (e.g., Antoniou et al., 2008) for longitudinal data analyses. In recent years, multilevel modeling and fractional regressions have also been employed as alternative research tools for the empirical analysis of capital structure (Bilgin & Dinc, 2019; Bilgin, 2019; Kayo & Kimura, 2011; Kieschnick & Moussawi, 2018; Ramalho et al., 2011).

This variety of research methods displays dissatisfaction with current approaches and a desire to search for better ones. Indeed, some criticisms are made for previous empirical corporate finance studies' model designs and estimation methods. The dynamic nature of capital structure restricts the explanatory power of static panel estimators and invalidates static approaches (Strebulaev, 2007; Strebulaev & Whited, 2013). Furthermore, the GMM-type estimators are also unreliable in the presence of endogeneity, unobserved heterogeneity, residual serial correlation, or changes in control parameters (Coles & Li, 2019; Dang et al., 2015; Flannery & Hankins, 2013; Grieser & Hadlock, 2019). Besides, nonlinearities should be taken into account in the model building since bounded dependent variables also result in biased estimates for linear specifications (Ashrafi, 2019; Elsas & Florysiak, 2015; Fattouh et al., 2005; Li et al., 2017; Ramalho et al., 2011).

Some researchers highlight econometric methods' shortcomings in finance research and claim that Machine Learning (ML) can be a better alternative to these methods, especially for high-dimensional data sets. As a state-of-the-art tool for understanding corporate decision-making, it recently gained popularity among researchers for analyzing investment analysis, asset valuation, and risk management decisions (Aziz et al., 2021). However, only a few studies use ML to investigate the financing choices of firms (Amini et al., 2021; Eliasy & Przychodzen, 2020; Sohrabi & Movaghari, 2020).

These techniques improve the model formation and estimation process in multiple aspects. First, when it is impossible to decide the correct functional form specification of the model, researchers have to base it on their prior knowledge or theoretical arguments, which are untestable (Belloni et al., 2014). Second, many econometric models should include only a restricted number

of control variables, or they may result in overfitting when the number of explanatory variables is close to the number of observations. Third, hypothesis testing is used to detect the significance of the model variables in econometrics, but the researchers arbitrarily determine the significance levels of these tests. Besides, different combinations of variables are tried to find a set with significant *p*-values (Simmons et al., 2011). This p-hacking behavior causes biased estimates due to the omitted variable bias, which restricts the predictive power of econometric models. On the other hand, OLS and other econometric methods overfit the noise when the number of explanatory variables gets closer to the number of observations (Nagel, 2021).

Econometric methods cannot solve the high-dimension problem of specifying a model with all relevant explanatory variables (Nagel, 2021). Fortunately, ML can solve these problems by reliably selecting the best control variables/instruments among a large set of possible ones by learning from the data and making predictions based on this training. Supervised learning methods of ML, which are used to predict a dependent variable (output) using several explanatory variables (features), are up-and-coming in corporate finance research (Amini et al., 2021; Feng et al., 2020; Rapach & Zhou, 2021; Sohrabi & Movaghari, 2020; Yang et al., 2020).

3 | DATA AND METHODOLOGY

3.1 | Data

The sample data set comprises 32,401 firms from 58 countries. Firm-specific data are extracted from Compustat Data Base. The number of sample firms for each country differs due to data availability. Table 1 presents the number of firms for sample countries. The sample period covers 18 years, from 2002 to 2019. Macroeconomic factors are collected from World Bank Data Base. In line with the general practice in capital structure research, we exclude financial firms and firms with negative shareholder's equity. Variables are winsorized to 1% and 99% to eliminate outliers.

The dependent variable of this study is the leverage ratio, estimated as the division of total long-term debt to the book value of total assets. The variables of interest are cash holdings, non-debt tax shield, and current ratio. Cash holdings are estimated as the ratio of cash and cash equivalents to the total assets and indicate the immediate liquidity of the firm. The current ratio is the ratio of current assets to current liabilities and is related to the firms' working capital management. Both cash holdings and current ratio are considered negative proxies of financial distress probability. Investors tend to overvalue firms with more cash holdings and higher current ratios since their liquidity suffices to fulfill short-term liabilities. The perceived financial soundness of these firms enables them to raise external finance from the stock markets. However, too much liquidity may indicate underinvestment in long-term assets and causes lower return rates for investors. In a recent study, Sun and Xia (2022) show that when firms expect debt financing will be more expensive or risky, they can secure their future debt financing by issuing long-term debt now and holding the proceedings in cash. Consequently, firms' liquidity and capital structure policies are interlinked.

Furthermore, the non-debt tax shield is expected to put downward pressure on the leverage ratio by substituting the tax-shield effect of debt financing (Zafar et al., 2019). On the other hand, firms with high non-debt tax shields indicate high amounts of fixed assets, which can be used as collaterals in debt covenants and secure cheap bank credits (Acedo-Ramírez & Ruiz-Cabestre, 2014; Jaworski & Czerwonka, 2019).

The control variables of this study comprise both long-established determinants of leverage in the literature and some potential factors. Huenermund et al. (2022) warn against using high dimensional settings with many covariates. Following their suggestion of using a selected set of variables with theoretical reasoning, this study uses profitability, firm size, asset tangibility, and R&D as the firm-specific control variables. In addition to these, three industry-specific variables (i.e., uniqueness, regulated industry, and high-tech industry) and four country-specific variables (i.e., bank credit, GDP growth, NWC financing, and institutional quality) are controlled in our analysis. Time and country dummies are also included to isolate time- and country-specific fixed effects. Variable definitions are given in Table A1 in the Appendix.

Table 2 gives the descriptive statistics of the variables. The average leverage ratio of the total sample is .13, which moves within the 12%–14% range over the sample period. This ratio increases to .30 for some countries and as low as .09 for others. The sample distribution of the leverage ratio is right-skewed and slightly peaked. While the overall average of cash holdings is .16, it varies within the .07–.26 range at the country level and .13–.18 over time. This variable's distribution also has a fat tail, but its peak is higher than the leverage ratio. Our second focus variable, the non-debt tax shield, has a sample average of .04. Its country-level range is .02–.06, and its time-based range is .03–.04. Its distribution is almost symmetric but highly peaked. Lastly, the current ratio has an almost time-invariant overall sample mean of 2.62 with a country-level range from 1.16 to 5.98. Its distribution is both skewed and highly peaked. The pairwise correlations of variables are presented in Table A2 in the Appendix.

TABLE 1 Number of sample firms for each country.

Country	Firms	Percent	Country	Firms	Percent
Argentina	76	.23	Korea Rep.	1944	6.00
Australia	2362	7.29	Kuwait	89	.27
Austria	94	.29	Luxembourg	60	.19
Bangladesh	196	.60	Malaysia	1036	3.20
Belgium	137	.42	Mexico	123	.38
Brazil	352	1.09	Netherlands	221	.68
Bulgaria	58	.18	New Zealand	189	.58
Cayman Islands	1175	3.63	Nigeria	105	.32
Chile	175	.54	Norway	309	.95
China	3709	11.45	Oman	70	.22
Colombia	37	.11	Pakistan	343	1.06
Croatia	83	.26	Philippines	181	.56
Cyprus	75	.23	Poland	742	2.29
Denmark	179	.55	Portugal	53	.16
Egypt	142	.44	Russia	286	.88
Finland	182	.56	Saudi Arabia	130	.40
France	893	2.76	Singapore	736	2.27
Germany	889	2.74	Slovenia	30	.09
Greece	242	.75	Spain	178	.55
Hong Kong	271	.84	Sri Lanka	196	.60
Hungary	37	.11	Sweden	815	2.52
India	3502	10.81	Switzerland	267	.82
Indonesia	459	1.42	Thailand	629	1.94
Ireland	97	.30	Tunisia	49	.15
Israel	419	1.29	Turkiye	321	.99
Italy	409	1.26	United Arab Emirates	48	.15
Japan	4054	12.51	United Kingdom	2312	7.14
Jordan	115	.35	Vietnam	482	1.49
Kenya	38	.12	**Total**	**32401**	**100.00**

Source: Author's computation.

3.2 | Methodology

The following partially linear regression model is estimated in this study:

$$\begin{aligned} y &= \delta g + f_0(H) + \varepsilon \; E(\varepsilon|g,H) = 0 \\ g &= m_0\;(H) + \vartheta \; E\;(\vartheta|H) = 0 \end{aligned} \tag{1}$$

where y is the dependent variable, g the variable of interest, H is the matrix of the control variables, and δ is the treatment effect. The functional forms with respect to H, (f_0, m_0) are unknown.

Partialing-out and double selection are the two solutions to obtain an estimate of δ in Equation (1) when the models are linear and sparse (Belloni et al., 2012, 2014). The double selection procedure of Belloni et al. (2014) provides a reliable inference on the coefficient of g guarding against the omitted variable bias due to model misspecification by selecting the control variables with a two-pass estimation process. Firstly, a LASSO regression of y is run on H. Next, a LASSO regression of g is run on H. This second LASSO regression reveals the factors not selected the first step but may cause omitted variable bias due to their correlation with g. Once the LASSO estimator selects the factors in the previous two steps, a cross-sectional OLS regression of y is run on the covariances selected in these two steps.

On the other hand, the partialing-out procedure of Belloni et al. (2012) can be used which requires multiple estimation steps. First, a LASSO regression of y is run on H and $\tilde{H}_y$, which is the set of selected covariates of this initial LASSO, is obtained. Second, a LASSO of y on $\tilde{H}_y$ is run and the residuals $(\tilde{y})$ are saved. Similarly, a LASSO regression of g on H is run and the selected covariates $(\tilde{H}_g)$ are

TABLE 2 Descriptive statistics of variables.

	Mean	Median	St.Dev.	Skew.	Kurt.	Min.	Max.	*N*
Dependent variable								
Leverage ratio	.13	.08	.16	1.42	4.89	.00	1.00	354,961
Variables of interest								
Cash holdings	.16	.11	.17	1.83	6.81	.00	1.00	354,961
Non-debt tax shield	.04	.03	.04	5.84	83.42	.00	.99	354,961
Current ratio	2.62	1.52	4.85	9.25	122.06	.00	99.80	354,961
Control variables								
Profitability	.07	.08	.14	−2.07	15.04	−1.00	1.00	354,961
Firm size	7.50	7.32	3.17	.28	2.84	−3.82	23.40	354,961
Asset tangibility	.31	.27	.23	.68	2.72	.00	1.00	354,961
Bank credit	2.95	2.27	3.28	.01	4.45	−15.15	24.00	354,961
NWC financing	3.05	2.14	4.20	2.60	21.43	−25.96	50.92	354,961
GDP growth	3.00	2.14	3.99	2.09	17.17	−25.96	50.92	354,961
Institutional quality	3.53	4.67	5.08	−.20	1.51	−7.59	11.82	354,961

Source: Author's computation.

obtained. Then, a LASSO of g on $\tilde{H}_g$ is run and the residuals ($\tilde{g}$) are saved. Lastly, an OLS regression of $\tilde{y}$ on $\tilde{g}$ is run. The slope coefficient of this last regression is the coefficient estimate (δ).

An increasingly popular alternative is the cross-fit partialing-out LASSO framework of Chernozhukov et al. (2018), which splits the data into k sub-samples and employs a version of the partialing-out algorithm. This approach is also known as the DML and enables the use of other machine learning algorithms.

In this study, we employ the double selection LASSO (Belloni et al., 2012), partialing-out LASSO (Belloni et al., 2014; Chernozhukov et al., 2015a; 2015b), and cross-fit partialing-out LASSO (Chernozhukov et al., 2018) procedures for the estimation of δ's assuming linear functional forms for (f_0, m_0). All these inference procedures account for the omitted variable bias and systematically evaluate the contribution of a new factor to the explanatory power of an existing model.

4 | FINDINGS

Tables 3 and 4 present the estimation results for the roles of cash holdings, non-debt tax shield, and current ratio as determinants of capital structure controlling the variables listed in Panel C of Table A1 in the Appendix and the country and year effects. Initially, the within transformation is used to eliminate the panel fixed effects of the data set, as the built-in commands of STATA 17 software for LASSO inference cannot be applied to panel models. The double-selection LASSO method of Belloni et al. (2014), the partialing-out LASSO method of Belloni et al. (2012) and Chernozhukov et al. (2015a, 2015b), and the cross-fit partialing-out LASSO estimator (DML) of Chernozhukov et al. (2018) are employed to analyze the transformed dataset using the STATA 17 commands; dsregress, poregress, and xporegress, respectively. Plugin and adaptive methods are used to select the tuning parameters. The estimated coefficients, standard errors, and the t-statistics of focus variables are given in Table 3. All three methods detect the significant and negative impact of cash holdings on the leverage ratio, which is robust to the tuning parameter selection method. Cash holdings have a downward pressure on the indebtedness of a firm.

The non-debt tax shield is found to have a positive effect on the leverage ratio. However, the effect becomes insignificant when the plugin method is used with the double selection or partialing-out estimators. Similarly, these two estimators fail to detect any significant relationship between the current ratio and firm leverage for both plugin and adaptive lambda selection procedures. However, the cross-fit partialing-out reveals that the current ratio has a downward pressure on the indebtedness.

For panel data analysis, pdslasso, a STATA command of Ahrens et al. (2018), enables the double-selection, partialing-out and cross-fit partialing-out methods for panel data structures used to estimate the coefficients for the variables of interest. The results are presented in Table 4. The findings align with the ones presented in Table 3, but the significant results are more prevalent than the pooled estimations. Both cash holdings and current ratio are found to have a highly significant positive effect on the leverage ratio, while the non-debt tax shield

TABLE 3 The estimation results based on LASSO inference methods with pooled data.

Selection method for Lambda:	Dependent variable: Leverage ratio					
	Plugin			Adaptive		
	estimate	s.e.	*t*-ratio	estimate	s.e.	*t*-ratio
Cash holdings						
Double-selection	−.1003	.0082	−12.2600	−.0927	.0091	−10.1600
Partialing-out	−.1013	.0075	−13.4300	−.0925	.0092	−10.0900
The cross-fit partialing-out	−.1012	.0031	−32.7000	−.0927	.0036	−25.8200
Non-debt tax shield						
Double-selection	.0151	.0092	1.6400	.0258	.0083	3.0900
Partialing-out	.0151	.0090	1.6800	.0257	.0081	3.1700
The cross-fit partialing-out	.0151	.0040	3.8000	.0255	.0035	7.2900
Current ratio						
Double-selection	−.0094	.0107	−.8800	−.0079	.0110	−.7100
Partialing-out	−.0090	.0107	−.8500	−.0079	.0110	−.7200
The cross-fit partialing-out	−.0091	.0037	−2.4700	−.0079	.0039	−2.0200

Notes: The table reports the results of the LASSO inference methods for the roles of cash holdings, non-debt tax shield, and current ratio as determinants of capital structure controlling the variables listed in Panel C in the Appendix and the country effects. The data set is split into $k = 10$ folds for the cross-fit partialing-out estimations. Plugin and adaptive methods are used for the selection of the tuning parameter. The standard errors are clustered by company and year. The estimates are obtained using three built-in commands of STATA 17 software (i.e., dsregress, poregress, and xporegress) after within transformation of the data set.
Source: Author's computation.

TABLE 4 The estimation results based on LASSO inference methods for panel data.

Selection method for Lambda:	Dependent variable: Leverage ratio		
	Plugin		
	Estimate	s.e.	*t*-ratio
Cash holdings			
Double-selection	−.0862	.0031	−28.0100
Partialing-out	−.0852	.0031	−27.5000
Cross-fit Partialing-out	−.1493	.0076	−19.5700
Non-debt tax shield			
Double-selection	.0990	.0143	6.9200
Partialing-out	.0966	.0144	6.7200
Cross-fit Partialing-out	.0578	.0119	4.8520
Current Ratio			
Double-selection	−.0002	.0001	−3.2300
Partialing-out	−.0002	.0001	−3.1000
Cross-fit Partialing-out	−.0503	.0067	−7.5460

Notes: The table reports the results of the LASSO inference methods for the roles of cash holdings, non-debt tax shield, and current ratio as determinants of capital structure controlling the variables listed in Panel C in the Appendix and the country effects. The data set is split into $k = 6$ folds for the cross-fit partialing-out estimations. Plugin and adaptive methods are used for the selection of the tuning parameter. The standard errors are clustered by company and year. The results are obtained using PDSLASSO, a user-written STATA command (Ahrens et al., 2018). Standard errors are adjusted for 32,401 clusters in firms.
Source: Author's computation.

decreases it. All three methods detect high significance for these relationships.

Our findings align with Sohrabi and Movaghari (2020), who also use the LASSO method and report that liquidity measures and non-debt tax shields are stable determinants of book-based leverage. In another recent study using the LASSO and non-linear machine learning algorithms, Amini et al. (2021) also suggest liquidity as a valid determinant of firm leverage.

5 | ROBUSTNESS TESTS

We have run several additional tests to check the robustness of the results against the possible effects of the Great Financial Crisis on the results. The sample is divided into two sub-samples covering 2002–2008 and 2009–2019, and models are re-estimated for each sub-sample separately. The pooled data results are presented in Table 5.

The sub-sample results for the pooled data are found to be very similar to the total sample results for cash holdings. Cash holdings' significant and negative impact on the leverage ratio is persistent in both sub-samples.

In line with the total sample results, the non-debt tax shield is also found to positively affect the leverage ratio during the 2009–2019 period. However, the same effect can be detected only when the cross-fit partialing-out estimator is employed with the adaptive method for the 2002–2008 period.

TABLE 5 The sub-sample results based on LASSO inference methods with pooled data.

Selection method for Lambda:	Subsample 2002–2008						Subsample 2009–2019					
	Dependent variable: Leverage ratio											
	Plugin			Adaptive			Plugin			Adaptive		
	Estimate	s.e.	*t*-ratio	Estimate	s.e.	*t*-ratio	estimate	s.e.	*t*-ratio	Estimate	s.e.	t-ratio
	Cash holdings											
Double-selection	−.0904	.0097	−9.3500	−.0862	.0113	−7.6000	−.0761	.0074	−10.3500	−.0767	.0072	−10.6200
Partialing−out	−.0904	.0097	−9.3700	−.0861	.0109	−7.9000	−.0757	.0075	−1.1400	−.0766	.0073	−1.4900
The cross−fit partialing−out	−.0904	.0048	−18.7000	−.0862	.0052	−16.7300	−.0757	.0035	−21.4900	−.0761	.0035	−22.0000
	Non−debt tax shield											
Double-selection	.0023	.0119	.2000	.0131	.0126	1.0400	.0419	.0084	4.9800	.0418	.0078	5.3300
Partialing−out	.0023	.0108	.2100	.0131	.0108	1.2100	.0419	.0085	4.9600	.0417	.0078	5.3900
The cross-fit partialing-out	.0023	.0053	.4400	.0132	.0057	2.3300	.0419	.0040	1.5000	.0422	.0039	1.7100
	Current ratio											
Double-selection	.0100	.0070	1.4300	.0105	.0081	1.3000	−.0099	.0100	−.9800	−.0099	.0103	−.9600
Partialing−out	.0100	.0067	1.4900	.0105	.0081	1.3000	−.0098	.0101	−.9700	−.0099	.0103	−.9600
The cross-fit partialing-out	.0100	.0039	2.5900	.0107	.0044	2.4300	−.0098	.0040	−2.4300	−.0102	.0039	−2.6000

Notes: The table reports the results of the LASSO inference methods for the roles of cash holdings, non-debt tax shield, and current ratio as determinants of capital structure controlling the variables listed in Panel C in the Appendix and the country effects. The data set is split into $k = 10$ folds for the cross-fit partialing-out estimations. Plugin and adaptive methods are used for the selection of the tuning parameter. The standard errors are clustered by company and year. The entire sample is divided into two sub-samples to cover the periods of 2002–2008 and 2009–2019. Then, models are estimated for each sub-sample separately. The estimates are obtained using three built-in commands of STATA 17 software (i.e., dsregress, poregress, and xporegress) after within transformation of the data set.
Source: Author's computation.

Similar to the total sample results, only the cross-fit partialing-out estimator detects the significant effect of the current ratio—however, the sign of the coefficient changes between the sub-samples. The current ratio has a downward pressure on the leverage ratio during 2002–2008. This finding is different from the total sample results. However, the relationship turns to be positive for the 2009–2019 period.

Table 6 gives the panel data estimation results for the sub-samples. Once again, the cash holdings preserve the previously detected significant and positive effect on the leverage ratios for both sub-sample periods. The positive and significant coefficient estimates of the non-debt tax shield also align with the total sample results for both sub-sample periods. The cross-fit partialling-out results for the current ratio are the same as this study's preliminary results and indicate a significant negative impact for both sub-periods. However, the two other methods detect a positive relationship between the current ratio and the leverage for the 2002–2008 period.

In sum, the robustness tests' findings support this study's main findings to a great extent. The most crucial difference is the sign of the effect of the current ratio for the pre-crisis period. However, this unexpected positive impact returns to negative when the cross-fit partialing-out estimator is employed in Table 6.

6 | CONCLUSION

Empirical capital structure literature recognizes a large set of firm-specific and macroeconomic factors as determinants of leverage. However, there is no consensus on the relative importance of these determinants since the empirical tests of any new factor rely on a randomly selected subset of the existing factors as control variables. The resulting models usually suffer from the omitted variable bias and have low explanatory powers. Besides, the prevalence of non-linear relationships and interactions between determinants of capital structure shadow the current research results and highlights the need for new methodological approaches to verify the existing knowledge and illuminate the unknown. Even though empirical studies on leverage have recently become more reliable due to advanced econometric methods, there is still a need to improve the robustness of their findings.

ML methods have recently become a popular analysis tool in finance research, which can help investigate firms' financing choices. This study contributes to the

TABLE 6 The estimation results based on LASSO inference methods for panel data.

	Subsample 2002–2008			Subsample 2009–2019		
Dependent variable: Leverage ratio						
Selection method for Lambda: Plugin						
	Estimate	**s.e.**	***t*-ratio**	**Estimate**	**s.e.**	***t*-ratio**
	Cash Holdings					
Double-selection	−.0860	.0052	−16.7000	−.0669	.0035	−19.0400
Partialing-out	−.0841	.0052	−16.2400	−.0665	.0035	−18.8100
Cross-fit Partialing-out	−.1490	.0064	−23.2050	−.1490	.0064	−23.2050
	Non-debt Tax Shield					
Double-selection	.0396	.0201	1.9800	.1678	.0185	9.0800
Partialing-out	.0403	.0202	1.9900	.1682	.0184	9.1200
Cross-fit Partialing-out	.0573	.0104	5.5000	.0573	.0104	5.5000
	Current Ratio					
Double-selection	.0002	.0001	2.0600	−.0003	.0001	−3.6500
Partialing-out	.0003	.0001	2.3400	−.0002	.0001	−3.3100
Cross-fit Partialing-out	−.0503	.0052	−9.6570	−.0504	.0052	−9.6570

Notes: The table reports the results of the LASSO inference methods for the roles of cash holdings, non-debt tax shield, and current ratio as determinants of capital structure controlling the variables listed in Panel C in the Appendix and the country effects. The data set is split into $k = 6$ folds for the cross-fit partialing-out estimations. Plugin and adaptive methods are used for the selection of the tuning parameter. The standard errors are clustered by company and year. The entire sample is divided into two sub-samples to cover the periods of 2002–2008 and 2009–2019. Then, models are estimated for each sub-sample separately. The results are obtained using PDSLASSO, a user-written STATA command (Ahrens et al., 2018). Standard errors are adjusted clusters in firms.
Source: Author's computation.

literature by advocating LASSO-based ML methods to decide the control variables for empirical capital structure studies. We to evaluate the marginal contributions of cash holdings, non-debt tax shield, and current ratio as determinants of capital structure. Our findings have revealed that they are crucial factors that substantially affect the leverage decisions of firms and should be controlled in empirical capital structure studies. To the authors' knowledge, the current study represents the first application of the LASSO inference methods in capital structure research.

Our approach is reliable for selecting the control variables for further empirical capital structure research by overcoming the omitted variable bias prevalent in the literature. It suggests a reliable method for investigating the effect of any new factor on the capital structure. Our findings provide a fresh perspective for future research on capital structure and corporate decision-making.

DATA AVAILABILITY STATEMENT

The author has provided the required Data Availability Statement, and if applicable, included functional and accurate links to said data therein.

ORCID

Rumeysa Bilgin https://orcid.org/0000-0002-5919-0035

REFERENCES

Acedo-Ramírez, M. A., & Ruiz-Cabestre, F. J. (2014). Determinants of capital structure: United Kingdom versus continental European countries. *Journal of International Financial Management & Accounting*, *25*(3), 237–270.

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2018). PDSLASSO and IVLASSO: Programs for post-selection and post-regularization OLS or IV estimation and inference. http://ideas.repec.org/c/boc/bocode/s458459.html

Alves, P. F. P., & Ferreira, M. A. (2011). Capital structure and law around the world. *Journal of Multinational Financial Management*, *21*(3), 119–150.

Amini, S., Elmore, R., Öztekin, Ö., & Strauss, J. (2021). Can machines learn capital structure dynamics? *Journal of Corporate Finance*, *70*, 102073.

Antoniou, A., Guney, Y., & Paudyal, K. (2008). The determinants of capital structure: Capital market-oriented versus bank-oriented institutions. *Journal of Financial and Quantitative Analysis*, *43*(1), 59–92.

Ashrafi, M. (2019). Nonlinear relationship between institutional investors' ownership and capital structure: Evidence from Iranian firms. *International Journal of Managerial and Financial Accounting*, *11*(1), 1–19.

Aziz, S., Dowling, M., Hammami, H., & Piepenbrink, A. (2021). Machine learning in finance: A topic modeling approach. *European Financial Management*, *2021*, 1–27.

Baum, C. F., Caglayan, M., & Rashid, A. (2017). Capital structure adjustments: Do macroeconomic and business risks matter? *Empirical Economics*, *53*(4), 1463–1502.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. B. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*, 2369–2429.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.

Bilgin, R. (2019). Relative importance of country and firm-specific determinants of capital structure: A multilevel approach. *Prague Economic Papers*, *28*(5), 499–515.

Bilgin, R., & Dinc, Y. (2019). Factoring as a determinant of capital structure for large firms: Theoretical and empirical analysis. *Borsa Istanbul Review*, *19*(3), 273–281.

Chakrabarti, A., & Chakrabarti, A. (2019). The capital structure puzzle–evidence from Indian energy sector. *International Journal of Energy Sector Management*, *13*(1), 2–23.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. B., Newey, W. K., & Robins, J. M. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, *21*(1), C1–C68.

Chernozhukov, V., Hansen, C. B., & Spindler, M. (2015a). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, *105*, 486–490.

Chernozhukov, V., Hansen, C. B., Spindler, M., & Partialing-out lasso linear regression 9. (2015b). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, *7*, 649–688.

Coles, J. L., & Li, Z. F. (2019) An empirical assessment of empirical corporate finance. Available at SSRN: https://ssrn.com/abstract=1787143

Czerwonka, L., & Jaworski, J. (2022). Capital structure and its determinants in companies originating from two opposite sides of the European Union: Poland and Portugal. *Economics and Business Review*, *8*(1), 24–49.

Dang, V. A., Kim, M., & Shin, Y. (2015). In search of robust methods for dynamic panel data models in empirical corporate finance. *Journal of Banking & Finance*, *53*, 84–98.

De Jong, A., Kabir, R., & Nguyen, T. T. (2008). Capital structure around the world: The roles of firm-and country-specific determinants. *Journal of Banking & Finance*, *32*(9), 1954–1969.

Eliasy, A., & Przychodzen, J. (2020). The role of AI in capital structure to enhance corporate funding strategies. *Array*, *6*, 100017.

Elsas, R., & Florysiak, D. (2015). Dynamic capital structure adjustment and the impact of fractional dependent variables. *Journal of Financial and Quantitative Analysis*, *50*(5), 1105–1133.

Fattouh, B., Scaramozzino, P., & Harris, L. (2005). Capital structure in South Korea: a quantile regression approach. *Journal of Development Economics*, *76*(1), 231–250.

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *75*(3), 1327–1370.

Flannery, M. J., & Hankins, K. W. (2013). Estimating dynamic panel models in corporate finance. *Journal of Corporate Finance*, *19*, 1–19.

Frank, M. Z., & Goyal, V. K. (2009). Capital structure decisions: Which factors are reliably important? *Financial Management*, *38*(1), 1–37.

Gogas, P., & Papadimitriou, T. (2021). Machine learning in economics and finance. *Computational Economics*, *57*(1), 1–4.

Gonzalez, V. M., & Gonzalez, F. (2008). Influence of bank concentration and institutions on capital structure: New international evidence. *Journal of Corporate Finance*, *14*(4), 363–375.

Grieser, W. D., & Hadlock, C. J. (2019). Panel-data estimation in finance: Testable assumptions and parameter (in) consistency. *Journal of Financial and Quantitative Analysis*, *54*(1), 1–29.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLos Biology*, *13*(3), e1002106.

Huenermund, P., Louw, B., & Caspi, I. (2022). Double machine learning and automated confounder selection—A cautionary tale. In Academy of Management Proceedings (Vol. *2022*, No. 1, p. 14311). Academy of Management.

Jaworski, J., & Czerwonka, L. (2019). Meta-study on relationship between macroeconomic and institutional environment and internal determinants of enterprises' capital structure. *Economic research-Ekonomska istraživanja*, *32*(1), 2614–2637.

Kayo, E. K., & Kimura, H. (2011). Hierarchical determinants of capital structure. *Journal of Banking & Finance*, *35*(2), 358–371.

Kieschnick, R., & Moussawi, R. (2018). Firm age, corporate governance, and capital structure choices. *Journal of Corporate Finance*, *48*, 597–614.

Li, T., Munir, Q., & Abd Karim, M. R. (2017). Nonlinear relationship between CEO power and capital structure: Evidence from China's listed SMEs. *International Review of Economics & Finance*, *47*, 1–21.

Lim, S. C., Macias, A. J., & Moeller, T. (2020). Intangible assets and capital structure. *Journal of Banking & Finance*, *118*, 105873.

Loncan, T. R., & Caldeira, J. F. (2014). Capital structure, cash holdings and firm value: A study of Brazilian listed firms. *Revista Contabilidade & Finanças*, *25*, 46–59.

Nagel, S. (2021). *Machine learning in asset pricing*. Princeton University Press.

Panda, A. K., Nanda, S., Hegde, A. A., & Yadav, A. K. K. (2021). Receptivity of capital structure with financial flexibility: A study on manufacturing firms. *International Journal of Finance & Economics*, *2021*, 1–13.

Pathak, R., Gupta, R. D., & Jalali, A. (2021). The analysis of debt levels in public firms: An international evidence. *Managerial Finance*, *47*(11), 1553–1570.

Poornima, B. G., & Kumar, P. (2022). A study on the capital structure determinants of FMCG companies in India. *International Journal of Financial Engineering*, *9*(02), 2150008.

Rajan, R. G., & Zingales, L. (1995). What do we know about capital structure? Some evidence from international data. *The Journal of Finance*, *50*(5), 1421–1460.

Ramalho, E. A., Ramalho, J. J. S., & Murteira, J. M. (2011). Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys*, *25*(1), 19–68.

Rapach, D., & Zhou, G. (2021). Asset pricing: Time-series predictability. Available at SSRN 3941499.

Rundo, F., Trenta, F., di Stallo, A. L., & Battiato, S. (2019). Machine learning for quantitative finance applications: A survey. *Applied Sciences*, *9*(24), 5574.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Sohrabi, N., & Movaghari, H. (2020). Reliable factors of capital structure: Stability selection approach. *The Quarterly Review of Economics and Finance*, *77*, 296–310.

Strebulaev, I. A. (2007). Do tests of capital structure theory mean what they say? *The Journal of Finance*, *62*(4), 1747–1787.

Strebulaev, I. A., & Whited, T. M. (2013). Dynamic corporate finance is useful: A comment on Welch. *Critical Finance Review*, *2012*(2), 173–191.

Sun, Q., & Xia, J. (2022). Cash holdings, capital structure, and financing risk. *Journal of Financial and Quantitative Analysis*, *57*(2), 790–824.

Yang, J. C., Chuang, H. C., & Kuan, C. M. (2020). Double machine learning with gradient boosting and its application to the Big N audit quality effect. *Journal of Econometrics*, *216*(1), 268–283.

Zafar, Q., Wongsurawat, W., & Camino, D. (2019). The determinants of leverage decisions: Evidence from Asian emerging markets. *Cogent Economics & Finance*, *7*, 1598836.

Zhu, D., Qiu, Z., & Wang, J. (2021). Factors affecting the capital structure of listed Chinese media companies. *International Journal of Finance & Economics*, *2021*, 1–10.

How to cite this article: Bilgin, R. (2023). The selection of control variables in capital structure research with machine learning. *Journal of Corporate Accounting & Finance*, *34,* 244–255. https://doi.org/10.1002/jcaf.22647

AUTHOR BIOGRAPHY

Rumeysa Bilgin has been an associate professor at Istanbul Sabahattin Zaim University since 2016. She received her BSc in Information Systems and Management from the University of London. She completed her Ph.D. in Finance at Istanbul University. Her research interests include corporate finance,and asset valuation. She is also interested in econometrics and machine learning methods used in empirical applications of finance theory.

Appendix

TABLE A1 Variable definitions and data sources.

Variables	Definition	Source
Panel A: The dependent variable		
Leverage ratio	The ratio of total long and short interest-bearing debt to total assets	Compustat
Panel B: Variables of interest		
Current ratio	The ratio of current assets to current liabilities	Compustat
Cash holdings	The ratio of cash and cash equivalents to total assets	
Non-debt tax shield	The ratio of depreciation expense to total assets	
Panel C: Control variables		
Profitability	The ratio of EBITDA to total assets	Compustat
Firm sze	Natural logarithm of total assets	
Asset tangibility	The ratio of plant, property and equipment to total assets *ppent/at*	
R&D	A dummy variable equals one if the firm reports a positive research and development expense, zero otherwise.	
Uniqueness	A dummy variable equals one if the firm's industry is producing sensitive products, zero otherwise. The classification is based on 4-digit SIC codes adopted by Amini et al. (2021).	
Regulated industry	A dummy variable equals one if the firm operates in regulated industries, zero otherwise. The classification is based on 4-digit SIC codes adopted by Amini et al. (2021).	
High-tech industry	A dummy variable equals one if the firm offers technology products and services, zero otherwise. The classification is based on 4-digit SIC codes adopted by Amini et al. (2021).	
Bank credit	Domestic Credit to Private Sector by Banks (%of GDP)—includes all credit to various sectors by deposit money banks, finance and leasing companies, money lenders, insurance corporations, pension funds, foreign exchange companies and other financial corporations.	World Development Indicators (World Bank)
GDP growth	GDP Growth Rate—Annual percentage growth rate of GDP at market prices based on constant local currency.	
NWC financing	Firms using banks to finance working capital- The percentage of firms using bank loans to finance working capital	
Institutional quality	The aggregate of six governance indicators (i.e., Control of Corruption, Government Effectiveness, Political Stability and Absence of Violence/Terrorism, Regulatory Quality, Rule of Law and Voice and Accountability.	World Governance Indicators (World Bank)

TABLE A2 Pairwise correlations of variables.

	Leverage ratio	Cash holdings	Non-debt tax shield	Current ratio	Profitability	Firm size	Asset tangibility	Bank credit	NWC financing	GDP growth	Institutional quality	R&D dummy	Uniqueness	Regulated industry	High tech industry
Leverage ratio	1.00														
Cash holdings	−.32	1.00													
Non-debt tax shield	.15	−.12	1.00												
Current ratio	−.18	.39	−.12	1.00											
Profitability	.04	−.16	.09	−.15	1.00										
Firm size	.14	−.13	−.06	−.16	.25	1.00									
Asset tangibility	.32	−.37	.18	−.14	.08	.14	1.00								
Bank credit	−.09	−.03	−.09	−.01	.09	.06	.06	1.00							
NWC financing	.04	−.12	−.02	.01	.08	−.02	.11	.26	1.00						
GDP growth	.04	−.12	−.02	.01	.08	−.03	.11	.27	.97	1.00					
Institutional quality	.04	.14	.11	.04	−.18	−.19	−.15	−.57	−.46	−.47	1.00				
R&D dummy	−.08	.12	.00	−.02	−.01	.19	−.16	−.10	−.21	−.21	.14	1.00			
Uniqueness	−.10	.05	−.01	−.01	.00	.04	−.13	.05	−.08	−.08	.01	.27	1.00		
Regulated industry	.14	−.06	.08	−.05	.05	.12	.16	−.01	.06	.06	−.03	−.06	−.10	1.00	
High tech industry	−.07	.15	.12	.03	−.02	−.07	−.18	−.04	−.07	−.07	.09	.17	.24	.06	1.00

Source: Author's computation.