Baruch College

Zicklin School of Business

ECO 9723

December 12, 2021

Title:

Detection of Earnings Manipulation using published data

Table of Contents

Abstract 3
Introduction 4
Data and variable description 6
Method of Data Analysis 7
Empirical Analysis 11
Conclusion 15
Appendix 17
References 20

Word count (Tables and Table descriptions have not been included):

Section	Word count
Introduction	243
Data and Sample Formation	322
Methods of Data Analysis	511
Empirical Analysis	832
Conclusion	213
Total	2121

Abstract

The paper investigates a large number of earnings manipulators and aims to examine their distinguishing characteristics. To achieve this, accounting line items of manipulators' and non-manipulators' financial statements has been analyzed and logistic regressions were used to develop models, which are able to predict if a firm is manipulating its financials. The regressors of the model have been created in such a way they may capture the effect of manipulation or some preconditions that induce the firm to engage in earnings management or accounting fraud. Moreover, the regressors employed in the paper were mentioned in previous research on this topic (Beneish, 1999; Dechow, 2011). The results of our models suggest that there is a relationship between accounting ratios and probability of manipulation.

Since earnings manipulation constitutes a problem, not only for auditors, but also for investors and other market participants, these models may be used to identify most of the companies involved in earnings manipulation before public discovery. Even though these models can be easily implemented, a firm may be individuated as a manipulator due to possible distortions in its financial statements, also if it has not altered its financials.

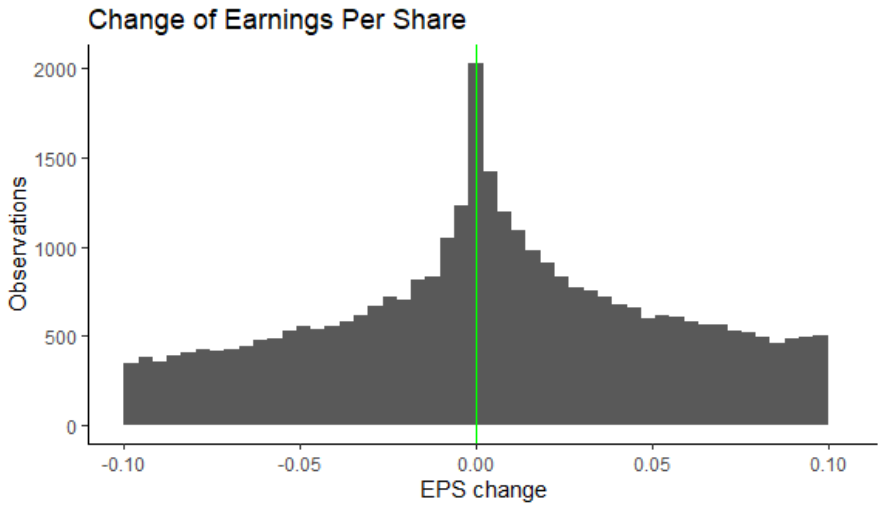
Introduction

Altering financial statements is a well-known practice in accounting. The main problem related to financial statement falsification is the misrepresentation of the true picture of the firm also called "Accounting fraud" (Reurink, 2018). As a matter of fact, the most common types of alterations involve the overstatement of revenues, the understatement of expenses and the capitalization of costs as assets, and they may cause a considerable damage to investors of fraudulent companies (e.g., Enron and Parmalat SPA). However, financial misstatements are difficult to detect. Figure 1 represents the two distributions of change of EPS (Earnings per Share) and change in ROA (Return on Assets) from one year to another. Although Earnings should reflect ROA since firm performances are the same, the two distributions are slightly different: the EPS distribution is negatively skewed indicating that companies tend to declare zero or positive EPS rather than negative EPS. The ROA distribution in Figure 1b) is instead more symmetrical. One possible explanation of this event is earnings management and misstatements of financial reports.

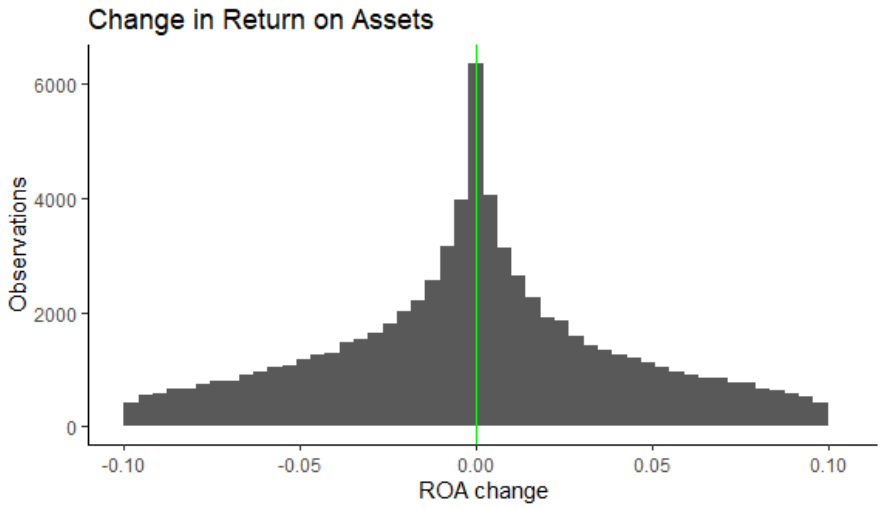
As a result, timely identification of fraudulent financial statements is important for market participants. Many models have been developed to detect fraudulent firms using ratios computed from financial statements and two of them (Beneish, 1999; Dechow et al. 2011) employed a logistic regression. The objective of this paper is to replicate the logistic regressions performed in these papers using recent data and potentially add new predictors to increase the performance of the model.

Figure 1: The database used to plot the following two graphs is the same that has been used in the development of the model. Figure a) shows change in Earnings per Share (EPS) which is calculated as the current EPS minus previous year EPS. Earnings per share (EPS) is defined as Net Income divided by Total number of shares outstanding. Figure b) shows Change in Return on Assets (ROA), which is calculated as current ROA minus previous year ROA. Return on Assets (ROA) is defined as Net Income divided by Total Assets.

a)



b)



Data and variable description

The database was provided by the research of Bao et al. (2019) and includes accounting data of quoted U.S. firms from 1990 to 2014 downloaded from COMPUSTAT. A comprehensive list of downloaded data can be found in Appendix 1. Furthermore, fraudulent firms that have been subject to enforcement actions by the Securities Exchange Commission (SEC) have been included after the examination of Accounting and Auditing Enforcement Releases (AAER). Generally, when an AAER is issued by the SEC, an investigation concluded that the firm has misstated its financial statements. Accordingly, the database considers a yearly observation as fraudulent if in that period the firm was accused of accounting fraud.

The use of AAER has advantages and disadvantages. Despite AAERs assure that the selected firms have manipulated their financial statements, not all fraudulent firms are likely to be investigated by the SEC. Indeed, the SEC may choose firms according to some red flags causing selection bias (Appendix 2).

After the cleaning process the number of observations in the database are 101,732 (Appendix 3). The number of manipulators of their financial statements were 311 out of 13,872 companies in the entire database and the total amount of misstated observations are 759. It is possible to notice that the number of manipulated and non-manipulated observations differ considerably resulting in an unbalanced database.

Table 1: Number of manipulators and non-manipulators. The number differs widely since firms that have been accused to have manipulated accounting data for at least one year are 2.24% of the total amount of firms in the dataset and altered financial statements are 0.75% of the total amount of observations.

	Manipulators	*Non-Manipulators*	*Total*
Companies	311	13,561	**13,872**
Observations (years)	759	100,973	**101,732**

Table 2 reports the industry distribution of misstating firms according to the SIC classification used in Dechow et al. (2011). It is evident that most of the manipulators were active in the “Durable Manufacturers” (26.69%) and “Computers” sectors (21.54%). Other sectors that were heavily influenced were the “Retail” (14.47%) and “Services” (12.86%) industries. The type of accounting item impacted by earnings management could vary depending on the industry. “Durable goods” and “Retail” are industries where misstatements are realized by capitalizing expenses as assets or create fictitious sales to meet the optimistic analysts’ forecasts. Contrarily, in the “Computers” industry, firms tend to overstate intangible assets.

Table 2: Distribution of manipulators by industry. The Standard Industrial Classification (SIC) code is the same as the classification code used in Dechow et al. (2011). Not all the companies in the database had a SIC and these were reported as "No Code" (NA).

Industry (SIC code)	*Misstating firms*	*Misstating firm (%)*	*Misstating obs.*	*Misstating obs. (%)*
Agriculture (0100-0999)	2	0.64	5	0.66
Mining and Construction (1000-1299, 1400-1999)	4	1.29	5	0.66
Food and Tobacco (2000-2141)	8	2.57	32	4.22
Textile and Apparel (2200-2399)	6	1.93	11	1.45
Lumber, Furniture, and Printing (2400-2796)	6	1.93	11	1.45
Chemicals (2800-2824, 2840-2899)	6	1.93	14	1.85
Refining and Extractive (1300-1399, 2900-2999)	6	1.93	18	2.37
Durable Manufacturers (3000-3569, 3580-3669, 3680-3999)	83	26.69	177	23.35
Computers (3570-3579, 3670-3679, 7370-7379)	67	21.54	206	27.18
Transportation (4000-4899)	8	2.57	28	3.69
Utilities (4900-4999)	6	1.93	16	2.11
Retail (5000-5999)	45	14.47	86	11.35
Services (7000-7369, 7380-9999)	40	12.86	55	7.26
Banks and Insurance (6000-6999)	11	3.54	31	4.09
Pharmaceuticals (2830-2836, 3829-3851)	11	3.54	20	2.6
No code (NA)	2	0.64	44	5.69
TOTAL	**311**	**100**	**759**	**100**

Method of Data Analysis

The dependent variable is binary (1=manipulator, 0=non-manipulator). Consequently, nonlinear regression models should be used because they force predicted values to a number between 0 and 1. Logistic regressions are considered to be the more appropriate for the purposes of this paper. This nonlinear regression model estimates the coefficients using the maximum likelihood estimator, which is normally distributed and consistent in large samples (Formula 1-2).

Formula 1: Probability of being a manipulator.

$$p(x) = P(Y = 1 \mid X = x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}}$$

Formula 2: Probability of being a non-manipulator.

$$p(x) = P(Y = 1 \mid X = x) = \frac{1}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}}$$

The most important variables available to market participants are made available by the SEC in form of financial data. Indeed, ratios or changes of these variables between years can be used as regressors in the model

and can be divided in three categories: Accruals quality, Financial performance and Market-based measures (Table 3).

Table 3: Regressors employed in the models. The column "Research" shows in which research the variable has been mentioned (Beneish (1999) = B, Dechow et al. (2011) = D, General ratio used in accounting = G).

Variable	Name in Database	Pred. sign	Research	Calculation
Misstate		-	B, D	Binary variable equal to 1 in case of misstatements and 0 otherwise
Accruals quality related variables				
RSST Accruals	ch_rsst	+	D	$\frac{(\Delta WC + \Delta NCO + \Delta FIN)}{Average\ total\ assets}$ *WC = Working capital* *NCO = Long-Term Assets - Long-Term Liabilities* *FIN = Investments - Debt + Preferred Stock*
Change in Receivables	dch_rec	+	D	$\frac{\Delta\ Accounts\ Receivable}{Average\ Total\ Assets}$
DSRI (Sales Index)	dsri	+	B	$\frac{Receivables_t/Sales_t}{Receivables_{t-1}/Sales_{t-1}}$
Change in Inventory	dch_inv	+	D	$\frac{\Delta\ Inventory}{Average\ Total\ Assets}$
% SOFT assets	soft_assets	-	D	$\frac{\left(Total\ Assets\ -\ Property, Plant\ and\ Equipment\ -\ Cash\ and\ Cash\ Equivalents\right)}{Total\ Assets}$
AQI (Asset Quality Index)	aqi	+	B	$\frac{1-(Current\ assets_t + PPE_t)/Total\ Assets_t}{1-(Current\ assets_{t-1} + PPE_{t-1})/Total\ Assets_{t-1}}$
DEPI (Depreciation Index)	depi	+	B	$\frac{Depreciation_{t-1}/(Depreciation_{t-1} + PPE_{t-1})}{Depreciation_t/(Depreciation_t + PPE_t)}$
TATA (Total Accrual)	tata	+	B	$\frac{\Delta\ Current\ Assets\ -\ \Delta\ Cash\ -\ \left(\Delta\ Current\ liabilities\ -\ \Delta\ Current\ long\ term\ debt\ -\ \Delta\ Income\ Tax\ Payable\right) - Depreciation\ and\ Amortization_t}{Total\ Asses}$
Measure of financial performance				
Change in cash sales	ch_cs	-	D	$\frac{Sales - \Delta\ Accounts\ Receivable}{Sales}$
Change in ROA	ch_roa	+	D	$\frac{Earnings_t}{Average\ Total\ Assets_t} - \frac{Earnings_{t-1}}{Average\ Total\ Assets_{t-1}}$
SGI (Sales Growth Index)	sgi	+	B	$\frac{Sales_t}{Sales_{t-1}}$

SGAI (Sales and General Administrative Expenses Index)	sgai	+	B	$\frac{Selling, General\ \&\ Administrative\ Exp._t/Sales_t}{Selling, General\ \&\ Administrative\ Exp._{t-1}/Sales_{t-1}}$
GMI (Gross Margin Index)	gmi	+	B	$\frac{(Sales_{t-1} - Cost\ of\ goods\ sold_{t-1})/Sales_{t-1}}{(Sales_t - Cost\ of\ goods\ sold_t)/Sales_t}$
Market-based measures				
Actual Issuance	issue	+	D	Binary variable (1 if the firm issued securities during year t, 0 if the firm did not issue any security)
Book-to-Market ratio	bm	+	G	$\frac{Book\ Value\ of\ Equity}{Market\ Value\ of\ Equity}$
LVGI (Leverage Index)	lvgi	-	B	$\frac{(LTD_t + Current\ liabilities_t)Total\ Assets_t}{(LTD_{t-1} + Current\ liabilities_{t-1})Total\ Assets_{t-1}}$

The first category, "Accrual quality", comprises the examination of accounting line items where the company makes assumptions. Firstly, "Change in Receivables", "Sales Index" and "Change in Inventory" always involve revenues or inventories, which could both highly depend on what firms consider as sales. Indeed, managers tend to overstate them and possibly sell on credit also growing the amount of their account receivables. Regarding inventory, managers tend to increase the level of inventory and fixed assets to sustain future growth, but when growth does not occur, they manipulate earnings by reducing depreciation (DEPI). Secondly, the SOFT Index measures the percentage of assets that are not Property Plant and Equipment or Cash, while the AQI Index determines how high Intangible Assets and Goodwill are. As mentioned before for the "Computers" segment, Intangible Assets and Goodwill are assets that are subject to assumptions and the more of these assets are included in the financial statement, the greater is the flexibility of the manager in managing earnings in the short-term. All these reasons are reflected in Table 4 where means and medians of manipulators are higher than means and medians of non-manipulators.

The second category of ratios involves "Financial Performance". Generally, fraudulent firms have poor performances and therefore, a decreasing Return on Assets, while non-manipulators are expected to perform better. Moreover, sales (SGI) are expected to increase more for manipulators than non-manipulators (Table 4) since some firms sell products to related companies or sell an enormous quantity of products at the end of the year with exaggerated return provisions to overstate revenues. Finally, manipulators would try to understate expenses to boost gross profit (GMI Index) or operating profit (SGAI Index).

The third category relates to "Market-based measures". According to economic theory, a large amount of debt will diminish the probability of misstatements of a firm because banks and debtholders will attentively control the firm performances. Because of this, Leverage (LVGI) has been included in the regression. Moreover, it has been proved that firms perform earnings management before or during the issuance of equity to show perfect financials and raise more capital. For this reason, Issuance was included as a binary variable that measures if the company has or not issued securities in the current period. Finally, the Book-to-Market ratio has been included to see if large companies (in terms of market value) are more likely to engage in fraud.

Table 4: Mean and Median for each regressor. Number of observations manipulators: 759. Number of observations non-manipulators: 100,973. It is possible to notice that the mean and median of misstated companies is generally higher than the mean and median of non-misstated companies for many predictors.

Variable	Name in Database	Mean (Misstated)	Mean (Non-Misstated)	Median (Misstated)	Median (Non-Misstated)
Accruals quality related variables					
RSST Accruals	ch_rsst	0.070	0.015	0.040	0.022
Change in Receivables	dch_rec	0.029	0.011	0.016	0.006
DSRI (Sales Index)	dsri	1.080	1.152	1.138	0.992
Change in Inventory	dch_inv	0.16	0.006	0.002	0.000
% SOFT assets	soft_assets	0.641	0.532	0.691	0.557
AQI (Asset Quality Index)	aqi	1.263	0.928	1.171	0.987
DEPI (Depreciation Index)	depi	1.138	1.123	1.136	1.010
TATA (Total Accrual)	tata	-0.074	-0.100	0.013	-0.041
Measure of financial performance					
Change in cash sales	ch_cs	0.273	0.184	0.138	0.072
Change in ROA	ch_roa	-0.013	-0.005	-0.005	-0.001
SGI (Sales Growth Index)	sgi	1.258	1.235	1.354	1.072
SGAI (Sales and General Administrative Expenses Index)	sgai	1.169	1.083	1.153	0.991
GMI (Gross Margin Index)	gmi	1.263	1.143	1.368	1.066
Market based measures					
Actual Issuance	issue	0.975	0.883	1	1
Book-to-Market ratio	bm	0.524	0.523	0.698	0.487
LVGI (Leverage Index)	lvgi	1.081	1.142	1.139	1.004

Empirical Analysis

Table 5 shows regression results of four models (out of 14 computed with the available regressors) and contains coefficients and heteroskedasticity robust standard errors.

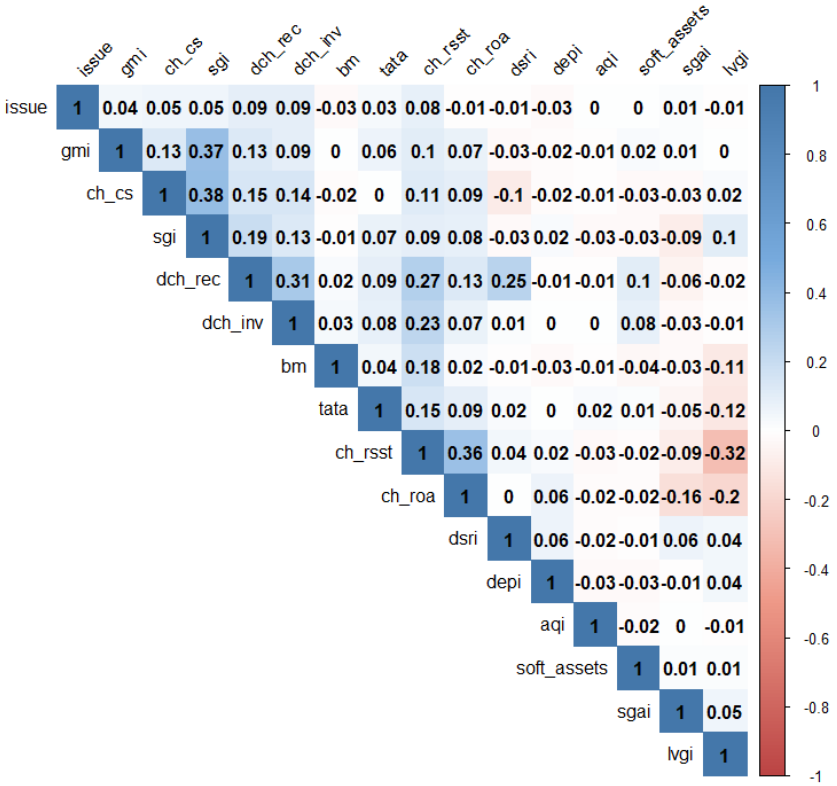
Model (1) includes all the regressors of the model developed by Beneish (1999). However, the results of the model differ slightly from the original paper since here LVGI is significative. Overall, the only factors that were considered significant at a 5% level were DSRI, SGAI, LVGI and GMI. According to this model, a decline in leverage seems to increase the probability of misstatements of financials, as well as a sudden increase in Revenues in comparison to Selling, General and Administrative costs (measured by SGAI). Furthermore, a significant GMI demonstrates that firms with a reduction in their gross margin, and a deterioration of their financial performances have a higher probability to engage in earnings manipulation. Differently from the original paper, the coefficient of DSRI is negative indicating a negative relationship between accruals and manipulators.

Model (2) reveals the results of the model developed by Dechow et al.(2011). All the regressors chosen in the model are significative (5%), as in the mentioned paper, except for Change in Inventory. It is possible to conclude that a significant increase in SOFT Assets or accrual variables ("Change in Receivables" and "Change in Sales") will improve the probability that the company is a manipulator. Concerning financial performance, a decrease in ROA will cause an increase in probability of manipulation because companies with declining performances are more likely to manipulate their financials. As mentioned before, an abnormal surge in sales will improve the probability of a firm being a manipulator. Finally, the coefficient of Issuance is positive and significative meaning that many manipulating firms issue securities when they engage in earnings management.

A considerable number of regressors have been developed from the same accounting data and to avoid multicollinearity the correlation between variables has been measured. Generally, all the correlation coefficients seem to be low due to transformations of line items in ratios (Figure 2). However, it has been considered appropriate to choose "Change in receivables" over DSRI and "Change in Cash Sales" over SGI

since in both cases the variables were calculated using the same elements. Moreover, both DSRI and SGI were less or not significative than "Change in receivables" and "Change is Cash sales".

Figure 2: Correlation plot. The correlation between SGI and "Change in Cash Sales" (ch_cs) is positive and amounts to 0.38 since both variables concern sales growth ("Change in Cash Sales" takes into consideration also the receivables and for this reason the correlation is not 1). Also the correlation between "Change in receivables" (dch_rec) and DSRI is positive and amounts to 0.25 due to the fact that receivables have been used to calculate the ratio.



Model (3) included all the variables of the two previously analyzed models. This model allows us to check the robustness of the results obtained in the previous regressions. It is possible to notice that only some Beneish coefficients changed slightly such as GMI, AQI, DEPI and TATA. For example, GMI is not significant anymore, while depreciation seems to have become significant at 10% level. Overall, all the other coefficients that were found significant in the previous two regressions remained significant without changing substantially. Similar tests have been conducted to check if results hold also when we subtract regressors.

The last model was developed with a stepwise backward and forward selection of regressors. Both procedures used the Akaike Information Criterion (AIC) to decide which regressor could be included in the model and both selected the same regressors. Moreover, this model is also the model that has the lowest AIC than any other model analyzed so far (Table 7) meaning that the level of Log Likelihood is high accounting for the

number of parameters included, and the model fits well with the data. All the regressors included in the model are significative at 1% except SGAI (5%) and LVGI (10%).

Lastly, the overall goodness of fit of the model has been analyzed by calculating the McFadden Pseudo-R-Squared, the Likelihood Ratio test and the Hosmer-Lemeshow statistic. First, since these are logit models, the usual R-squared may not be used evaluate their explanatory power. Consequently, the McFadden Pseudo-R-Squared was calculated, but typically, this measure has lower values than the R-squared in linear regression: the highest value of Pseudo-R-Squared is the value of Model 3 (3.25%) followed by model 4 (3.21%) indicating a low explanatory power of both models. Second, the Hosmer-Lemeshow null hypothesis is rejected at 5% significance level only in the first model indicating the difficulty of model 1 in predicting the outcome. Finally, since model (4) $\subset$ model (3), it is possible to test using the LR test the joint null hypothesis on the logit coefficients that there is a reduced model. The statistic fails to reject the hypothesis (5% significance) that model (4) (the reduced model) is better. For this reason, both models can be considered as appropriate.

Another important remark concerns internal and external validity threats. The possibility of sample selection bias has already been discussed (Appendix 2). Other threats may arise are error in variables and omitted variable bias. First, error in variables may occur when copying data published in the companies' financial statements leading to excessively high or low ratios. Second, other unobserved variables such as the quality of management or the level of compliance may be relevant and correlated with one regressor causing omitted variable bias.

Table 5: Results of four different regression models. The level of significance of each coefficient is denoted by asterisks: *p<0.1; **p<0.05; ***p<0.01.

	BENEISH (1)	DECHOW ET AL. (2)	FULL MODEL (3)	FORWARD & BACKWARD SELECTION (4)
DSRI	−0.087**			
	(0.038)			
GMI	0.051***		0.021	
	(0.015)		(0.015)	
AQI	−0.003		0.005	
	(0.018)		(0.017)	
SGI	−0.007			
	(0.023)			
DEPI	0.027		0.041*	
	(0.034)		(0.031)	
SGAI	0.039***		0.039**	0.040**
	(0.014)		(0.017)	(0.017)
LVGI	−0.126**		−0.121**	−0.111*
	(0.058)		(0.061)	(0.060)
TATA	0.033		−0.010	
	(0.040)		(0.030)	
CH RSST		0.569***	0.480***	0.515***
		(0.100)	(0.106)	(0.103)
DCH _REC		1.122**	1.201***	1.381***
		(0.436)	(0.436)	(0.423)
DCH _INV		0.747	0.799	
		(0.565)	(0.570)	
SOFT ASSETS		1.867***	1.878***	1.890***
		(0.160)	(0.160)	(0.159)
CH CS		0.052***	0.054***	0.063***
		(0.018)	(0.018)	(0.018)
CH ROA		−0.349***	−0.355***	−0.344***
		(0.084)	(0.083)	(0.083)
ISSUE		1.569***	1.570***	1.575***
		(0.233)	(0.233)	(0.233)
BM			0.0003	
			(0.020)	
CONSTANT	−4.776***	−7.533***	−7.529***	−7.471***
	(0.099)	(0.250)	(0.262)	(0.255)

Table 6: Chi squared and p-Values Testing Exclusion of Groups of Variables. Tests have been conducted to see the significance of groups of variables and specifically, the significance of the three categories: Accrual quality, Financial performance and Market-Based measures. Financial performance ratios, leverage and issue of securities all play an important role in the prediction of fraudulent firm (p-values lower than 0.01). However, coefficients of "Accrual ratios" as formulated by Beneish (1999) are not considered significative at any reasonable level meaning that accruals in that model are not relevant.

	BENEISH (1)	DECHOW ET AL. (2)	FULL MODEL (3)	FORWARD SELECTION (4)
Accrual quality	6.3389 (0.1752)	206.67 (<0.001)	202.6 (<0.001)	200.53 (<0.001)
Financial performance	18.574 (<0.001)	24.328 (<0.001)	36.061 (<0.001)	36.67 (<0.001)
Market Based measures	4.8252 (0.0281)	45.465 (<0.001)	48.838 (<0.001)	48.495 (<0.001)

Table 7: Statistics that measure the fit of the regression: Akaike Information Criterion, Pseudo-R-Squared developed by McFadden, Hosmer-Lemeshow statistic and Likelihood-Ratio statistic.

	BENEISH (1)	DECHOW ET AL. (2)	FULL MODEL (3)	FORWARD SELCTION (4)
Observations	101,732	101,732	101,732	101,732
Log Likelihood	-4,465.095	-4,332.286	-4,328.247	-4,330.038
Akaike Inf. Crit.	8,948.190	8,680.572	8,686.493	8,678.076
Pseudo R squared (McFadden)	0.00195	0.03164	0.03254	0.03214
Hosmer Lemeshow statistic (p-value)	21.033 (0.00759)	14.049 (0.08049)	13.819 (0.08661)	7.3783 (0.4964)
LR test statistic			3.583 (0.7329)	

Conclusion

This paper analyzed an adequate number of firms investigated by the SEC because involved in financial reports misstatements. Initially, the report focused on industries where most misstatements happened and then the characteristics of misstating firms on various dimensions such as accruals, financial performance, and other market-related variables. The results of two important papers on this topic (Beneish 1999, and Dechow et al. 2011) have been tested and other models have been developed by selecting appropriate ratios. Furthermore, limitations of all the models mentioned and identified in this paper have been described. In particular, other

than possible omitted variable bias and error in variables, selection bias may arise since misstatements go often undetected.

The main aim of the paper was to develop a model, which would provide insights to auditors, investors, or regulators that a firm is engaging in earnings management and financials manipulation. In particular, it is widely recognized that financials constitute an important method to deliver information to capital market participants and should represent the true picture of the firm financial situation. Future research may try to investigate, if the firms recognized by the developed models as fraudulent have manipulated their financial statements or their reports were legitimate and try to increase the explanatory power of the model by adding more regressors.

Appendix

Appendix 1: List of variables included in the dataset. Source: Walker, 2021.

Position	Column	Description
1	fyear	Fiscal Year
2	gkvey	Compustat firm identifier
3	sich	4-digit Standard Industrial Classification Code (SIC)
4	insbnk	An indicator variable for financial institutions between SIC 6000–6999
5	understatement	An indicator variable if the misstate indicator involved an understatement
6	option	Not used
7	p_aaer	Identifier for AAER
8	new_p_aaer	New Identifier for AAER
9	misstate	Indicator variable for misstatement
10	act	Current Assets - Total
11	ap	Accounts Payable - Trade
12	at	Assets - Total
13	ceq	Common/Ordinary Equity - Total
14	che	Cash and Short-Term Investments
15	cogs	Cost of Goods Sold
16	csho	Common Shares Outstanding
17	dlc	Debt in Current Liabilities
18	dltis	Long-Term Debt Issuance
19	dltt	Long-Term Debt Total
20	dp	Depreciation and Amortization
21	ib	Income Before Extraordinary Items
22	invt	Inventories - Total
23	ivao	Investment and Advances Other
24	ivst	Short-Term Investments - Total
25	lct	Current Liabilities - Total
26	lt	Liabilities - Total
27	ni	Net Income (Loss)
28	ppegt	Property, Plant and Equipment - Total (Gross)
29	pstk	Preferred/Preference Stock (Capital) - Total
30	re	Retained Earnings
31	rect	Receivables Total
32	sale	Sales/Turnover (Net)
33	sstk	Sale of Common and Preferred Stock
34	txp	Income Taxes Payable
35	txt	Income Taxes - Total
36	xint	Interest and Related Expense - Total
37	prcc_f	Price Close - Annual - Fiscal

Appendix 2: Selection bias

Another small remark should be considered: the SEC may or may not find all the firms guilty of fraudulent actions and misstated financial statements. One advantage of using AAER issued by the SEC is that firms either have been investigated and found guilty or have publicly admitted having misstated their financial

statements. However, it is possible that fraudulent firms might be unidentified. Consequently, this would reduce the predictive ability of a model built on accounting ratios. Moreover, a second disadvantage that may arise is selection bias: the SEC may decide to investigate a firm only after a suspect stock reaction or some red flags. According to authors in this field, selection bias is a general concern when investigating fraudulent firms and the determinants of earnings manipulation and does not happen only with the analysis of AAER (Dechow et al., 2011).

Appendix 3: Cleaning data process.

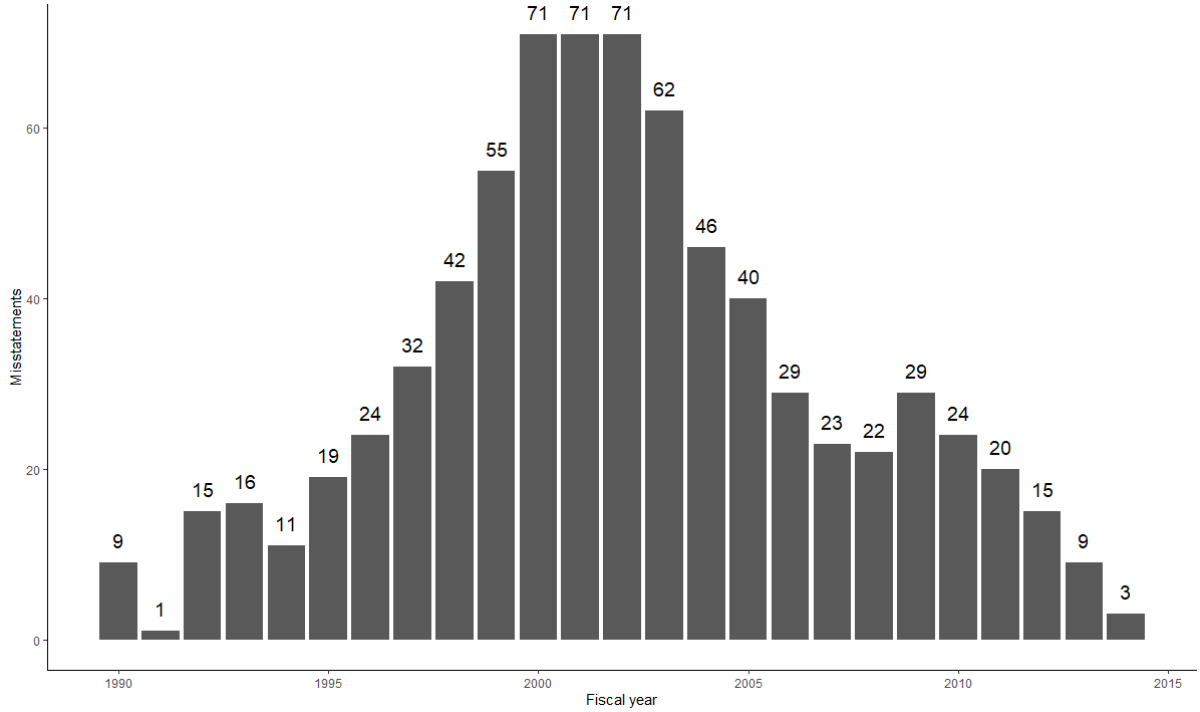
Unfortunately, the database contains some errors, which could affect the results of the models and for this reason, the data was cleaned to remove "NAs" and "Inf" errors in R. In particular, if either Revenues, Account receivables, and Total Assets were negative, the observation was dropped since these values cannot be negative and were necessary in the computation of some ratios. The same process has been applied if there was missing data for Revenues, Account receivables, and Total Assets because these accounting metrics were used as denominators in many ratios potentially causing "NAs".

After having calculated the ratios used in the regression model, all observations containing one or more "NA" or one or more "Inf" values have been dropped due to possible errors that may occur in fitting the model. Furthermore, it is necessary to mention that NAs have been added by one function that was created. The function added "NA" in each first-year (t) ratio available for that company if for the computation of that ratio data (t-1) was required and previous year data was not available. The initial database had 146,045 observations, while the cleaned database had just 106,402 observations.

Some data regarding the regressors obtained during this process were extremely high, which may then cause high-leverage influential points. For this reason, a cutoff value has been used, in order to improve the overall fitting of the model. A reasonable cutoff value chosen cutoff was [+20, -20] for the calculated ratios. As a result, the total number of observations remained in the dataset are 101,533.

Figure A3 shows the distribution of AAER for each fiscal year. The sample selected in this paper cover misstatements occurred between 1990 and 2014. A considerable number of misstatements occurred between 2000 and 2003 due to an increase an increase in the number of technology companies, a sector relatively new that was not extensively regulated providing further incentives to earnings management.

Figure A3: Number of AAER issued each year by the Securities Exchange Commission directed toward firms.



References

Bao, Y., Ke, B. I., Li B. I., Yu, Y. J., & Zhang, J. I. (2020). Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach. *Journal of Accounting Research*, 58(1), 199–235. https://doi.org/10.1111/1475-679X.12292

Beneish, M. D. (1999). The Detection of Earnings Manipulation. *Financial Analysts Journal*, 55(5), 24–36. https://doi.org/10.2469/faj.v55.n5.2296

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting Material Accounting Misstatements *Contemporary Accounting Research*, 28(1), 17–82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

Reurink, A. (2018). Financial Fraud: A Literature Review. *Journal of Economic Surveys*, 32(5), 1292–1325. https://doi.org/10.1111/joes.12294

Walker, S. (2021). Critique of an Article on Machine Learning in the Detection of Accounting Fraud. *Econ Journal Watch Scholarly Comments on Academic Economics,* 18(1), 61-70.