

## Repositório ISCTE-IUL

---

**Deposited in *Repositório ISCTE-IUL*:**

2026-08-28

**Deposited version:**

Accepted Version

**Peer-review status of attached file:**

Peer-reviewed

**Citation for published item:**

Trigueiros, D. (2019). Improving the effectiveness of predictors in accounting-based models. *Journal of Applied Accounting Research*. 20 (2), 207-226

**Further information on publisher's website:**

10.1108/JAAR-01-2018-0006

**Publisher's copyright statement:**

This is the peer reviewed version of the following article: Trigueiros, D. (2019). Improving the effectiveness of predictors in accounting-based models. *Journal of Applied Accounting Research*. 20 (2), 207-226, which has been published in final form at <https://dx.doi.org/10.1108/JAAR-01-2018-0006>. This article may be used for non-commercial purposes in accordance with the Publisher's Terms and Conditions for self-archiving.

---

### Use policy

Creative Commons CC BY 4.0

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in the Repository
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

---

# Improving the effectiveness of predictors in accounting-based models\*

Duarte Trigueiros

ISTAR-IUL, University Institute of Lisbon

## Abstract

Financial ratios are used as predictors in modelling tasks that require accounting information. This paper explains how the effectiveness of these models can be enhanced. Firstly, it discusses the inadequacies in the use of ratios as predictors. It then presents a methodology for constructing accounting-based models that overcome these inadequacies. Assuming plausible distributions of accounting data, the paper demonstrates that the modelling algorithm can eliminate the effect of size while identifying the optimal set of ratios for the most effective modelling. Empirical comparisons involving bankruptcy prediction, financial misstatement detection, and earnings forecasting confirm that the new methodology outperforms the conventional approach, eliminating the need for ratio selection and making model construction less arbitrary without compromising interpretability.

**Keywords:** Financial Ratio, Ratio Selection, Accounting-based Model, Predictive Modelling, Size Control, Bankruptcy Prediction, Financial Misstatement Detection, Earnings Forecasting.

**JEL codes:** C18, C21, C51, C58, M41.

## I. Introduction

Financial institutions, analysts, and investors, increasingly rely on multivariate modelling to assess credit risk, to estimate the likelihood of default, to forecast earnings' changes, to detect financial misstatement and to carry out other predictive tasks (Ak *et al.*, 2013). The purpose of this paper is to improve the effectiveness of accounting-based models, something that is probably a minor concern for research-driven modelling, but which is vital when the aim is to

---

\* Published: Trigueiros, D. (2019). Improving the effectiveness of predictors in accounting-based models, *Journal of Applied Accounting Research*, Vol. 20, No. 2, pp. 207-226. DOI: 10.1108/JAAR-01-2018-0006.

support actual decisions. In this paper, improved effectiveness means greater accuracy and explained variability (Wilcox, 2010).

Among factors determining model effectiveness, the paper attends to the adequacy of predictors. Accounting-based models use ratios as predictors. Ratios are indispensable in the analysis of financial statements where they bring to light financial features such as solvency or profitability while allowing the comparison of varied-size firms. But the fact that ratios are adequate in the context of financial analysis should not exempt them from scrutiny when the purpose is predictive modelling. The specific inadequacies of ratios when used as predictors are the need to find appropriate ratios among many alternatives (Tian *et al.*, 2015) and their erratic, statistically not well-behaved distributions (So, 1987; Lau *et al.*, 1995) that may lead to misrepresentation of essential qualities of the data being modelled (Kotz *et al.*, 2005).

The paper establishes a theoretical basis to demonstrate that the effect of size, which is removed by ratios, can also be removed by the modelling algorithm. While removing size, the algorithm portrays the financial features required by the modelled relationship as though the ratios were present while avoiding the above-mentioned inadequacies of ratios. A model-building methodology derived from this theoretical basis is proposed, capable of retaining ratios' most valuable trait, namely interpretability. Experiments are carried out to compare the new and the conventional methodologies. They cover well-documented classification tasks that illustrate low, medium and high difficulty levels, namely bankruptcy prediction, financial misstatement detection and earnings forecasting. The experiments verify that the new methodology delivers markedly more accurate, unbiased, and robust prediction.

The findings reported here have the potential to expedite the use of accounting-based models in predictive tasks where to date accuracy is not conclusive enough to support actual decisions. Better investment decisions, for instance, may result from a superior accuracy in predicting earnings' changes (Holthausen and Larcker, 1992). Also, in the face of concerns

about the declining value-relevance of accounting information (Lev and Gu, 2016), the gains in performance now at hand support the claim that such decline is not intrinsic to accounting information but may result from inadequately constructed models (Clout *et al.*, 2016).

## **II. Critical review of the use of ratios in accounting-based models**

This section exposes ratios' inadequacies when used as predictors in multivariate models, presenting the case for the proposed methodology and summarising related research. First, the difficulty presented by the need to seek appropriate ratios is outlined. Next, the section calls to mind the cause of ratios' erratic distributions and the ensuing consequences for modelling. Then, ratio-induced class misrepresentation is discussed in relation to current modelling practice. Finally, the propensity of ratios to bias models by selectively generating missing values is examined.

### ***The ratio selection process is unsatisfactory***

The selection of multivariate predictors is generally seen as a difficult task. Given the wealth of alternatives available, such difficulty increases when the predictors are financial ratios.

To solve this problem, some authors select predictors based on theory while other authors follow *ad-hoc* approaches. When the aim is to test hypotheses, ratios are pre-selected according to theory. But in Palepu (1986), Barnes (1990), Beneish (1999), and other studies where model effectiveness is a concern, ratios are equally selected based on individual qualities stated by theory. These studies are, in fact, transposing to the multivariate domain an essentially analytical, univariate routine. The consequence is that the predicting algorithm is limited in its choices and is unable to use all information available in sets of accounts.

Multivariate algorithms are synthesisers. They estimate moments and covariance matrices from multivariate distributions and then they extract from such estimations, and not from any specific ratio, whatever variability is optimal in modelling a relationship. This is

why models where effectiveness is important tend to favour the use of *ad-hoc* approaches to select ratios from a larger set (Altman, 1968; Du Jardin, 2009; Dechow *et al.*, 2011; Xu *et al.*, 2014; Tian *et al.*, 2015), with the view to obtain an optimal or nearly optimal prediction. But since the number of ratios that are potentially useful to model a relationship can be very large, some useful ratios may end up not being included in the selection procedure.

Both theory-driven and *ad-hoc* ratio selection are therefore unsatisfactory for applied modelling. In addition, the task of selecting ratios typically is time-consuming and laden with arbitrary choices. In Altman and Sabato (2007), for instance, this procedure involves four lengthy and not well-defined steps: theory- and literature-based ratio gathering, univariate pre-selection, multivariate selection and backward removal of weak predictors.

The methodology proposed in the paper bypasses ratio selection entirely. Variability required to model a relationship is identified as contained in the set of account items directly taken from financial statements, not in ratios. Rather than being at the inception of the model-building process, the finding of the optimal set of ratios is one of the results of this process.

### ***The statistical distribution of ratios may distort models***

The statistically not well-behaved distributions found in ratios have led authors interested in accounting-based models to propose ad hoc remedies such as winsorisation or outlier pruning (So, 1987; Lau *et al.*, 1995), ratio transformation (Watson, 1990; Altman and Sabato, 2007), or the use of robust estimation (Ohlson, 1980; Zmijewski, 1984). Today, these ad hoc remedies remain as popular as ever (Sun *et al.*, 2014; Ohlson and Kim, 2015; Leone *et al.*, 2017), suggesting that ratio distributions are still treated as oddities.

In fact, distributions of ratios are not odd, being easy to understand. Randomness, in continuous numbers, is the result of many interacting effects. Effects are aptly called additive when their likelihoods add up, or multiplicative when they multiply. Particularly, an additive distribution, such as the normal distribution, reveals the activity of additive effects whose

probability of occurrence is governed by the additive law of probability. A multiplicative distribution, such as the lognormal distribution, reveals the activity of multiplicative effects whose probability of occurrence is governed by the multiplicative law of probability.

It has long been established that income, wealth, firm size and other economic accruals resulting from the accumulation of random amounts obey a multiplicative law of probability, such as the Pareto or Gibrat laws, rather than an additive law, such as the central limit theorem (Ijiri and Simon, 1964; Singh and Whittington, 1968). Because the statistical behaviour of financial statement numbers approximates other economic accruals, their cross-sectional distributions obey a multiplicative law of probability (Tippett, 1990; Trigueiros, 1995; Falta and Willett, 2011), and the ratios formed from them are multiplicative (Johnson *et al.*, 1994). Building on this premise, McLeay and Trigueiros (2002) outlined the various types of distribution that can be found in ratios, and how to optimise them in each instance.

Multiplicative distributions are far from being well-behaved. In additive data, it is rare to find cases that differ from the mean by more than three standard deviations. For example, the height of an adult human is limited. Indeed, humans are neither taller than mountains nor shorter than mice. For multiplicative data, however, such disparities are not only possible, but likely. In fact, multiplicative distributions behave exponentially, i.e. when observations are ranked by size, they increase exponentially. Consequently, as the value of a multiplicative variable increases, the dispersion of cases increases very rapidly and their density decreases abruptly. In the resulting distribution, almost all cases are concentrated in a small interval, and a few cases are widely scattered (Aitchison and Brown, 1957; Johnson *et al.*, 1994).

The term ‘heteroscedasticity’, i.e. an uneven distribution of cases, does not do justice to the exponential behaviour of ratios. When it is claimed that Ordinary Least Squares (OLS) or Maximum Likelihood (ML) estimation, which are heteroscedasticity-inefficient but not biased, can be used in ratio-based modelling (Ohlson, 1980), it is clear that such a claim

ignores the multiplicative nature of ratios, where extremely influential cases will directly bias model coefficients, while the aforementioned abrupt drop in case density, which cannot be approximated by an OLS or ML fit (Kotz *et al.*, 2005), leads to models in which distributions are grossly misrepresented. It is also clear that extreme observations found in multiplicative data are not necessarily anomalous. Therefore, pruning or winsorisation is counterproductive in this case, being a source of bias (Leone *et al.*, 2017).

The methodology proposed in the paper is based on loglinear modelling (Aitchison and Brown, 1957), ensuring the accurate representation of multiplicative distributions, thus eliminating the need for ad hoc remedies. Eisenbeis (1977) misleadingly claimed that this model biases the value and sign of the coefficients. However, the loglinear function is ‘monotonically increasing’, implying that differences in sign are similar whether values are transformed or not. Therefore, covariance with other variables – and thus model coefficients – can differ only in magnitude, not sign. Although loglinear modelling is not suitable for testing the significance level of the original variables (Dambolena *et al.*, 2009), there is nothing to prevent its use in predictive modelling.

### ***Ratios may misrepresent class likelihoods in classification tasks***

Most accounting-based predictive models are classifiers (Amania and Fadlallah, 2017). One long-standing example (Altman, 1968, 1983) predicts firm failure using a ‘balanced’ design, whereby the number of failed and non-failed cases in each class is fairly equal. However, this balanced sampling and design procedure has fallen out of favour and is no longer used today, as the research community has been led to believe that it is inadequate.

At the root of this misunderstanding lies the notion that balanced designs result in a biased likelihood estimate (Palepu, 1986). However, when the purpose is prediction, this bias is irrelevant because it does not alter the relative positions of the likelihoods. Therefore, when the aim is to identify the cases most and least likely to belong to the class of interest, a

balanced design is not only unbiased, but also the most efficient way of representing the necessary information for modelling (Cosslett et al., 1981).

To avoid this alleged bias, the class frequencies in the datasets that are currently being employed for model building are made to reflect the underlying ‘prevalence’, i.e. the genuine a priori class probabilities. For instance, if the a priori probability of firm failure is 6 in 100, the model-building dataset should contain six per cent failed firms and ninety-four per cent non-failed firms. This approach inevitably results in models that are unusable because, when the prevalence is low, there will not be enough cases in the minority class to accurately represent its distribution (Zmijewski, 1984). This is particularly problematic in tasks such as predicting takeover targets (Palepu, 1986) or financial misstatements (Beneish, 1999), where both prevalence and predictive power are low. Furthermore, the class frequencies in the data made available rarely reflect prevalence. Therefore, the data must be manipulated before it is used, inevitably introducing a bias. Besides, if no attempt is made to address the exponential nature of the data, the predicted class frequencies will be misrepresented due to influential values, and it is pointless to make model-building class frequencies resemble prevalence.

Procedures are available that allow balanced sampling to be used for modelling while appropriately accounting for a priori class imbalance. Apart from the option of incorporating the class imbalance into estimation directly (Manski and MacFadden, 1981), it has been demonstrated that, in logistic regression, class imbalance only affects the intercept term. Consequently, a model can be developed using strictly balanced samples, and the intercept term can subsequently be adjusted to address the imbalance (Xie and Manski, 1989).

As the methodology proposed here puts an end to distribution misrepresentation in accounting-based models, efficient sampling designs can be used that do not require data manipulation. Where appropriate, a balanced design can be employed once more, paving the way for methods such as matching that effectively control for undesirable traits in the data.

### ***Ratio-generated missing values may also misrepresent class likelihoods***

Accounting-based empirical research consistently relies on third-party databases of financial statements, and there is an increasing concern that data quality issues may affect research findings. Casey *et al.* (2016) review the literature on the quality of databases of financial statements. Among other problems, these authors identify the presence of missing values. The use of ratios is also a source of missing values because a missing ratio component or a zero denominator creates a missing value. Missing values can distort models if the number of withdrawals is related to the variable being predicted. For example, companies that did not pay dividends in the previous period are improperly dropped from earnings predicting models where a popular predictor is the change in dividends per share from the previous period.

This paper presents clear instances of ratio-generated, class-dependent missing values, showing how the new methodology helps reducing the consequent bias.

### **III. A model-building methodology**

This section develops a theoretical foundation from which a two-stage model building methodology is derived. The underlying premise is that, except in well-known cases, financial ratios can be validly used in the analysis of financial statements. More than one hundred years of testing and intensive use supports this premise.

Ratios are widely employed in the financial analysis of firms because they eliminate the effect of size, thus enabling financial features to be compared across firms of different sizes. According to McLeay and Trigueiros (2002), ratios remove the effect of size because it is present in the two components of the ratio, which possess statistical distributions that are multiplicative — that is to say, effects multiply rather than add to each other. Indeed, when a quotient is formed from two numbers with a common divisor, the divisor is cancelled out.

The most general instance of a multiplicative behaviour is the lognormal distribution. When lognormality is hypothesised, the implications just mentioned require that the cross-

sectional variability of  $x_{ij}$ , i.e. the  $i^{\text{th}}$  item from the  $j^{\text{th}}$  set of accounts, be described as

$$\text{Log } x_{ij} = m_i + z_j + \epsilon_{ij} \quad (1)$$

where the logarithm of  $x_{ij}$  is explained as an expectation  $m_i$  accounting for the overall mean and the deviation from the mean specific to item  $i$ , plus  $z_j$ , the size effect specific to the set of accounts  $j$ , plus the size-unrelated unexplained variability  $\epsilon_{ij}$ . Formulation (1) is not intended to minutely portray the cross-sectional behaviour of items but to describe  $z_j$ , the common effect of size. The  $\epsilon_{ij}$  are not necessarily independent and, after size is accounted for, there may remain positive or negative correlations between the  $\epsilon_{ij}$  of any two items. According to (1), the logarithm of the ratio of items  $x_1$  and  $x_2$  taken from the same set of accounts is

$$\text{Log } \frac{x_2}{x_1} = \text{Log } x_2 - \text{Log } x_1 = (m_2 - m_1) + (\epsilon_2 - \epsilon_1) \quad (2)$$

where the effect of size,  $z_j$ , is no longer present. The median ratio  $\exp(m_2 - m_1)$  a suitable ‘ratio norm’ against which comparisons are made, while  $\exp(\epsilon_2 - \epsilon_1)$  is the ‘ratio feature’, i.e. the deviation from the norm observed in a ratio. For example,  $\exp(\epsilon_{\text{Expenses}} - \epsilon_{\text{Sales}})$  shows whether the Operating Ratio is above or below the norm. Thus,  $\epsilon_2 - \epsilon_1$  is, in logs, the information conveyed by a ratio about a specific financial feature of a specific firm.

If ratios are to be used as analytical tools, the implications leading to (1) must be verified. Otherwise, the ratios for large and small firms would behave differently. Solvency and profitability, for instance, would be an ineffective metric in financial analysis.

### ***Items where negative values may occur***

Formulation (1) does not apply to negative item values, such as those found in Net Income. However, account items where negative values may occur are subtractions of two positive-only items. Therefore, if  $x = x_A - x_B$  with positive  $x_A$  and  $x_B$ , then, for any sign of  $x$ ,

$$\text{Log } |x| = \text{Log } x_A + \text{Log } |1 - x_B/x_A|. \quad (3)$$

Since  $x_A$  is positive, (1) applies to  $\text{Log } x_A$ . And since, as shown in (2),  $x_B/x_A$  is size-

independent,  $z_j$  is not present in  $\text{Log}|1 - x_B/x_A|$ . Using the notation adopted in (1), the expected value of  $\text{Log } |x|$  is written as  $m'$ , being the result of adding  $m_A$  to the expected value of  $\text{Log}|1 - x_B/x_A|$ . Similarly, unexplained  $\text{Log } |x|$  are written  $\epsilon'$ . From (3),

$$\text{Log } |x| = m' + z_j + \epsilon' \quad (4)$$

for positive- and negative-valued  $x$ . Although  $m'$  and  $\epsilon'$  in (4) no longer have the meaning of  $m$  and  $\epsilon$  in (1), notably  $z_j$  in (4) is the same as in (1). This is the reason why ratios remove the effect of size, even when one of the components is negative.

Now, if ratios remove size no matter the sign of a component, it should be possible to use any item, not just positive-valued items, in modelling tasks requiring the removal of size. When items are transformed so that the effect of size is preserved regardless of sign, as in (4), their subtraction will remove size in a way similar to (2), also regardless of sign. One such size-preserving transformation is  $\text{Log } |x|$  coupled with a dummy retaining the sign of  $x$ . In classification tasks, dummies may be replaced by a fixed displacement  $d_{neg}$  of  $\text{Log } |x|$  for negative  $x$  only, where  $x$  is transformed into  $\text{Log } |x| - d_{neg}$ . Another transformation worth mentioning is  $\text{sign } x \text{ Log } |x|$ , known as Log-Modulus (John and Draper, 1980), where negative  $x$  are transformed into  $-\text{Log } |x|$ . Here, the sign of  $x$  is preserved, but a subtraction of two transformed items will not remove size from values that are negative. Therefore, Log-Modulus is only an alternative to consider in modelling tasks in which size is not an issue.

Zero-valued numbers should be treated as not multiplicative and kept as zero, because multiplicative distributions presuppose the existence of an accumulation. To note also that the mentioned transformations are equivocal within the range  $-1 < x < 1$  and cannot be used..

### ***Finding optimal predictors when the relationship is size-independent***

Consider the ‘loglinear’ relationship (Aitchison and Brown, 1957),

$$y = a + b_1 \text{Log } x_1 + b_2 \text{Log } x_2 + \dots \quad (5)$$

where  $x_1, x_2 \dots$  are item values taken from the same set of accounts and  $y$  is the attribute to be predicted. In the case of a classifier,  $y$  is a linear score. Where predictors obey (1), then (5) is

$$y = A + b_1\epsilon_1 + b_2\epsilon_2 + \dots + (b_1 + b_2 + \dots) z_j \quad (6)$$

where  $A = a + b_1m_1 + b_2m_2 + \dots$  is constant. That is, variability available to model  $y$  has a size-independent term  $b_1\epsilon_1 + b_2\epsilon_2 + \dots$  and a size-related term  $(b_1 + b_2 + \dots) z_j$ . If  $y$  is size-independent, the term  $(b_1 + b_2 + \dots) z_j$  must be zero to bar size-related variability from entering the relationship. Coefficients  $b_1, b_2, \dots$  must therefore add to zero. For instance, if two log-transformed item values are employed in the prediction of a size-independent  $y$ , then (5) becomes  $y = a + b_1 \text{Log } x_1 + b_2 \text{Log } x_2$  with  $b_2 = -b_1 = b$ , that is,

$$y = a + b(\text{Log } x_2 - \text{Log } x_1) \quad \text{or} \quad y = a + b \text{Log } \frac{x_2}{x_1}. \quad (7)$$

Ratio  $x_2$  to  $x_1$  is therefore formed in response to  $y$ 's size-independence. Supposing that the ratio  $x_2$  to  $x_1$  predicts bankruptcy accurately, when the logarithms of its components are shown to a modelling algorithm, the linear combination explaining bankruptcy optimally is found when  $b_2 = -b_1$  as in (7) because, at that point, size is removed, the ratio is formed, its predictive power is released. Similarly, when three items are employed in the prediction of a size-independent  $y$ , then  $b_3 = -b_1 - b_2$  and (5) now is  $y = a + b_1 \text{Log } \frac{x_1}{x_3} + b_2 \text{Log } \frac{x_2}{x_3}$ . That is, in response to size-independence, two ratios are formed from three account items.

In the general case, a size-independent  $y$  could indeed induce  $N - 1$  ratios from  $N$  items. The lost degree of freedom arises from the fact that size is no longer being modelled. However, algorithms cannot pair items two-by-two without external assistance. Rather, as part of the modelling process, some size-related variability would be assigned to a number of predictors to compensate for the size-related variability assigned to other predictors. This means that the algorithm would assign the role of 'ratio denominator' to some predictors (negative-signed  $b$  coefficients) and that of 'ratio numerator' to others (positive-signed  $b$

coefficients) so  $b_1 + b_2 + \dots = 0$ , the overall effect being the removal of size.

The resulting linear combination is a high-dimensional, size-independent log-ratio that can represent the financial features required by the relationship. To the extent that the transformations applied to item values are intended to preserve the effect of size, items with negative values will also be able to form linear combinations capable of removing size.

When, for example,  $\text{Log } |x| - d_{neg}$  is applied to  $x_1, x_2 \dots$  in (5), the term  $(b_1 + b_2 + \dots)(z_j - d_{neg})$  is formed, where  $d_{neg}$  is distinct, but constant for positive and negative values. Therefore, in this case as well, size-independent  $y$  may cause  $b_1 + b_2 + \dots = 0$ .

The conclusion to be drawn is that the task of selecting ratios capable of representing financial features while removing size is not a modelling prerequisite. In response to the size independence of a relationship, the algorithm will form optimal high-dimensional ratios from log-transformed items. In the case of two predictors only, the algorithm will form one ratio.

### ***The methodology***

As shown in equation (5), size-related  $y$  can be modelled directly by individual log-items, leading to optimal and well-behaved relationships. However, in this case, high-dimensional ratios would not be formed. Furthermore, unlike models that use ratios, any data instability would necessitate rebuilding the model. Models that use untransformed or transformed ratios plus a size proxy as predictors would be immune to this issue, but finding an optimum set of ratios is difficult for the reasons noted above.

Size-independent  $y$  can be modelled directly using log-transformed items, as shown in equation (5). In this case, financial features would emerge and the effect of size would be offset as if ratios were present. However, apart from adding to zero, the coefficients of the high-dimensional log ratios formed could not be interpreted, which conflicts with analysts' requirements. To achieve model interpretability while preserving the ability to identify the set of well-behaved, optimal predictors, the proposed methodology comprises two stages. In the

first stage, a loglinear model is built as shown in equation (5), with the sole purpose of selecting, from the comprehensive set of  $M$  log-items, an optimal subset  $N \ll M$  possessing predictive power. Then, in the second stage, pairs e.g.  $\text{Log } x_2 - \text{Log } x_1$  of the  $N$  predictors are subtracted. Like ratios, these pairs are size-independent and interpretable. No more than  $N - 1$  pairs should be formed and none of the  $N$  items should be omitted. The algorithm will employ these pairs as predictors to estimate the model with a degree of freedom less.

By selecting predictors from the comprehensive set of  $M$  items and not from the much more numerous and ill-defined set of potentially useful ratios, the new methodology greatly simplifies the predictor-selection process, while curbing the chances of overlooking variables that have predictive power. Practical suggestions for implementing the methodology follow:

1. Databases of financial statements typically scale all numbers down by one million ( $10^6$ ). Item values must be descaled to avoid falling within the range  $-1 < x < 1$ .
2. Transformation  $\text{Log } |x| - d_{neg}$  is suited for negative-valued items in classification.  $d_{neg}$  should exceed the descaling power, that is, for a  $10^6$  descaling,  $d_{neg} = 7$ .
3. In case the second-stage accuracy measure is smaller than the first-stage accuracy, the methodology is failing to deliver the optimal set of predictors, probably due to size-dependency. It may still be used, with a size-proxy, as a non-optimal alternative.
4. Changes  $\Delta x$  and  $\Delta x/x$  in relation to the previous period are both transformed into

$$\Delta \text{Log } x = \text{Log } x_t - \text{Log } x_{t-1} \quad (8)$$

where  $t - 1$  and  $t$  are subsequent periods. For positive and negative  $x$ ,  $\Delta \text{Log } x$  are size-related but, contrary to ratios, (8) will not create missing values when  $x_{t-1} = 0$ .

#### IV. The experiments

This section demonstrates how the new methodology is used and compares ratio-based models with models based on the new methodology. Comparisons cover bankruptcy prediction, financial misstatement detection and earnings forecasting, being carried out using

controlled experimental designs in which model characteristics, i.e. data used to build and test the models, class frequencies, modelling algorithm, randomisation and information provided, are strictly identical for the ratio-based model and the model based on the new methodology.

Experiments incorporate *ex-ante* and *ex-post* prediction, size-independent and size-related relationships and balanced, as well as unbalanced classes. Cases are not pruned, even in ratio-based models. Table 1 shows items to be tested for inclusion in the first stage of the new methodology. Before testing, item  $x$  is descaled into the original value and transformed:

- for  $x > 1$ ,  $x \rightarrow \text{Log } x$ .
- for  $x = 0$ ,  $x$  is not transformed.
- for  $x < -1$ ,  $x \rightarrow \text{Log } |x| - d_{neg}$  for bankruptcy prediction, and  $-\text{Log } |x|$  for earnings forecasting. No negative values occur in the misstatement detection experiment.

The specific transformation applied in the first stage is also used in (8). Logistic regression is the algorithm used throughout. First-stage predictor selection is performed via an ordinary variable-selection algorithm (e.g. ‘stepwise’). Multicollinearity is tested in the first stage.

-- Table 1 --

### ***Bankruptcy prediction***

This experiment reproduces Altman’s (1968, 1983) bankruptcy prediction models. Ohlson (1980), Zmijewski (1984) and Taffler (1982) discuss bankruptcy prediction and offer alternatives to Altman’s models. A review of this research, where high accuracy levels are customary, can be found in Balcaen and Ooghe (2006) and in Sun *et al.* (2014).

The experiment aims, not so much at illustrating the benefits of the new methodology, but at introducing its distinctive characteristics. The UCLA-LoPucki data<sup>1</sup> and other sources provide a total of 2,997 bankruptcy filings. Filings that are not the first to occur in each firm, and filings about which financial statements are not available, are discarded. Both ‘Chapter 7’

---

<sup>1</sup> <http://lopucki.law.ucla.edu/>

and ‘Chapter 11’ filings are included.<sup>2</sup> Two samples of nearly 900 cases each are drawn from the remaining filings. They contain firms that are present in the Compustat file and are listed in U. S. exchanges. Bankruptcy dates cover the period 1979-2008. All the ten size categories, computed as the deciles of the logarithm of Assets, and all the twenty-four Global Industry Classification Standard (GICS) groups are represented.

To reduce unwanted variability, firms in the two samples are matched with firms not bankrupt. Matching is based on the GICS group and on size tertile. Once a bankrupt case is assigned to its peer, this peer is made unavailable for matching. One of the datasets resulting from this matching is utilised to build the models and the other to measure their effectiveness.

The ratio-based model uses Altman’s (1983) X1 to X5 ratios, detailed in Table 2. As for the new methodology, the algorithm selects at its first stage five predictors from the X1 to X5 ratio components only. Negative values are transformed into  $\log |x| - d_{neg}$ , with  $d_{neg} = 7$  to exceed the  $10^6$  descaling.

Table 2 shows that accuracy is high in both models, but Chi-square and R-Square values are visibly higher in the first-stage model than in the model based on ratios. Therefore, ratios fail to use variability available in their components. Table 2 also compares the class frequencies in the model-building dataset with the class frequencies predicted by the model.

-- Table 2 --

Size-offsetting, mentioned in relation to (5), is visible in first-stage coefficients. Working Capital and Total Liabilities have positive signs and add to 9.74. These are the numerators of internal ratios. The denominators are Total Assets, Retained Earnings and Operating Income after Depreciation, with negative signs and adding to 10.35.

In the second stage, the five predictors selected in the first stage are paired and the components of each pair are subtracted, forming four predictors. Among other possible

---

<sup>2</sup> Under Chapter 11, debts are restructured; under Chapter 7, assets are liquidated to repay debts.

combinations, the algorithm estimates the coefficients leading to the score

$$0.95 + 0.13 \text{ Log } \frac{WCAP}{AT} + 7.92 \text{ Log } \frac{LT}{AT} - 0.47 \text{ Log } \frac{RE}{LT} - 0.43 \text{ Log } \frac{OIADP}{LT}$$

where abbreviations refer to Table 1 and, in order to emphasise model interpretability, subtractions of the two paired items are depicted as ‘Log’ followed by the fraction sign.

$\text{Log } \frac{WCAP}{AT}$ , for instance, really means  $(\text{Log}|WCAP| - d_{neg}) - \text{Log } AT$  where  $d_{neg}$  is zero for positive  $WCAP$  and 7 for negative  $WCAP$ . Zero  $WCAP$  remain zero.

This model has the same R-Square as that in the first stage. Chi-square is 1,912 with 4 degrees of freedom (df). Overall accuracy is 96 per cent with 97 per cent not bankrupt and 96 per cent bankrupt cases correctly classified. Performance is the same for other possible pairings of the 5 items. Since second-stage performance, which is based on size-independent predictors, equals first-stage performance, it is concluded that size is not required in this case.

### ***Detecting misstated sets of accounts***

Reported sets of accounts are extremely efficient in revealing financial features of firms. This efficiency is the ultimate reason why accounts are misstated so often (Eilifsen and Messier, 2000). Predictive models can help detect misstatement (Beneish, 1999; Dechow *et al.*, 2011; Sun *et al.*, 2014), typically attaining an overall accuracy of nearly 80 per cent for non-homogeneous samples and 93 per cent for homogeneous samples (Liou, 2008). However, in the models made public so far, the minority class is invariably poorly predicted, while the opposite class is very well predicted. This disparity renders models useless.

The second experiment compares the *ex-ante* effectiveness of a misstatement detection model based on ratios with that of a model based on positive-only, log-transformed items. Misstated accounts are defined as having been the object of an Accounting and Auditing Enforcement Release resulting from investigations made by the U. S. Securities and Exchange Commission for alleged accounting misconduct. Releases are divided in two sets,

one with cases from years 1978 to 1999 and one with cases from years 2000 to 2008. The 1978-1999 set is utilised in model-building, the 2000-2008 set is utilised to measure the effectiveness of models.<sup>3</sup> The two sets are matched with an equal number of firms which are neither the object of releases throughout the period nor are bankrupt in the release year. Matching is based on the GICS group and size tertile. Once a misstatement case is assigned to its peer, this peer is made unavailable for future matching.

The ratio-based model uses ratios selected from Beneish (1999) and other studies. As detailed in Table 3, twelve ratios are selected, five of them from Beneish. Accuracy is 74 per cent. False-positives are far more likely than false-negatives. The predicted class frequencies diverge widely from the class frequencies observed the model-building dataset.

-- Table 3 --

The dataset utilised to build the ratio model has 358 missing and 781 not missing cases. To allow comparison, these 358 missing cases are also excluded from the first- and second-stage modelling (although, in these, the number of missing cases would be 18 only). The items made available to the selection algorithm in the first stage of the new methodology are components of the ratios that entered the ratio-based model (and their changes, as in 8).

Table 3 shows the selected seven items and two changes. Overall accuracy is now 84 per cent and predicted class frequencies closely follow frequencies in the model-building dataset. The variability explained is almost twice that of the ratio-based model. It is therefore clear that ratios fail to use much variability available in their components. Size offsetting is visible: Cash, Liabilities and Capital, have negative-signed coefficients, magnitudes adding to 15.4. Assets, Receivables, Debt and Revenue, have positive signs, magnitudes adding to 16.2.

From the seven items selected in the first stage, only five pairs are formed, as nothing is to be gained by adding a sixth pair.  $\Delta \text{Log } LT$  is included, but the other change also

---

<sup>3</sup> Data from the Centre for Financial Reporting, Haas School of Business, University of California.

selected,  $\Delta \text{Log } CHE$ , is non-significant here. From such predictors, the score

$$\begin{aligned} -5.90 + 3.48 \Delta \text{Log } LT - 1.87 \text{Log } \frac{CHE}{REVT} - 9.75 \text{Log } \frac{AT}{LT} + 0.77 \text{Log } \frac{CSTK}{RECT} \\ + 0.26 \text{Log } \frac{CSTK}{DLTT} + 0.51 \text{Log } \frac{CSTK}{REVT} \end{aligned}$$

is estimated, where subtractions of the two transformed items are depicted as ‘Log’ followed by the fraction sign, and where ‘ $\Delta$ ’ refers to (8). The R-Square of this second-stage model is the same as that in the first stage. The Chi-square is 696 with 6 df. Overall accuracy is 84 per cent with 87 per cent not misstated and 81 per cent misstated cases correctly classified.

### ***Forecasting the direction of earnings’ changes one year ahead***

The third experiment is inspired by Ou (1990), who verified the ability of ratios to forecast the direction of unexpected changes in Earnings per Share one year ahead. Ou and Penman (1989), Holthausen and Larcker (1992) and others, also used Ou’s model in the prediction of market returns, while Bird *et al.* (2001), while reviewing the published research, examined the robustness of Ou’s model to different time periods and industries. The predictive gains are typically poor, by some six to ten per cent above the fifty per cent ‘chance’ level.

The distinctive features of the experiment are the absence of sampling, the very large number of cases (from which predicted classes are computed) and a size-related relationship. The experiment is also an example of the bias that ratio-generated, class-dependent missing values may create on the predicted class frequencies. Indeed, among the ratios used by Ou, the percentage change in Inventories to Assets has 20 per cent missing decreases in Earnings per Share, while for increases it is 43 per cent. Changes in Dividends per Share relative to the previous year have 29 per cent missing decreases, while for increases it is 57 per cent.

After excluding missing values, some 140,000 cases remain, where nearly 65 per cent are not increases and 35 per cent are increases. From this dataset, two samples with the same size are drawn at random, one for model-building and one for measuring the effectiveness of

models. All ratios in Ou (1990) plus other ratios are tested for inclusion. Ratios formed from Net Income or related items are not excluded from selection. The algorithm selects nine ratios as detailed in Table 4, of which three are from Ou. The logarithm of Assets is also included in the ratio-based model as a size proxy, bringing the total number of predictors to ten.

When building the model according to the new methodology, the predictors are taken exclusively from the ratio components selected by the ratio-based model and their changes, computed as in (8). The change in Dividends per Share relative to the previous year fails ratio-selection but, after being transformed as in (8), is selected to enter in the first stage. Negative values are transformed into  $-Log |x|$ . This would be inadequate for size-independent relationships but is not particularly damaging here. The first stage selects six items, plus Dividends per Share, plus three changes. As usual, Table 4 shows class frequencies, performance, plus the coefficients and their significance levels.

-- Table 4 --

The accuracy of the ratio-based model is 74 per cent with 89 per cent non increases and 47 per cent increases correctly classified. Class frequencies are grossly misrepresented in relation to the model-building dataset. By contrast, the first-stage accuracy is 80 per cent with no class misrepresentation. The variability explained is twice that obtained by ratios.

At the second stage, four pairs are formed from five of the six items selected in the first stage. The spare item,  $Log AT$ , takes the role of size proxy and is not paired. The score

$$0.420 - 0.108 Log \frac{RE}{LT} - 0.044 Log \frac{OANCF}{LT} - 0.061 Log \frac{TXT}{LT} - 0.095 Log \frac{GP}{LT} - 0.256 Log AT \\ + 0.036 \Delta Log RE + 0.054 \Delta Log GP + 0.107 \Delta Log DVPS - 0.205 Log DVPS$$

is estimated, where subtractions of the two transformed items are depicted as ‘ $Log$ ’ followed by the fraction sign, and ‘ $\Delta$ ’ refers to (8). The accuracy, balance and the R-Square of this second-stage model are the same as that at the first stage. The Chi-square is 27,143 for 9 df.

Figure 1 compares the ROC curves of the ratio-based and second-stage models

presented earlier, demonstrating the asymmetry of the former and the latter's superiority. The three experiments show that ratio-based models cannot utilise all of the available variability for modelling purposes, resulting in an inaccurate representation of the distributions. The improvements brought about by the new methodology are largely due to the predicted class frequencies resembling the model-building frequencies rather than to pure performance gains. Overall, the evidence from the three experiments should be sufficient to end the current uncritical use of ratios and unlock the full potential of accounting information in supporting value-creating decisions.

-- Figure 1 --

## **V. Final remarks**

The use of financial ratios as predictors in modelling tasks where accounting information is required is fraught with difficulties due to the need to select ratios among many alternatives and the multiplicative nature of the statistical distributions involved, which create influential values and may lead to misrepresentation of essential qualities of the data. To avoid these inadequacies, the paper has developed a theoretical foundation from which a two-stage modelling methodology was derived. The first stage of this methodology identifies the reduced set of transformed items that can optimally explain a relationship. The second stage uses pairs of the identified items to model the relationship. In addition to efficiency gains, the predictors used in modelling are now selected from a well-defined set of candidates and ad hoc remedies, such as the pruning of extreme cases, are no longer necessary. The three experiments documented in the paper demonstrate that the new methodology avoids the need to make arbitrary decisions while delivering more accurate, robust and interpretable models.

The scope of the methodology is limited. Firstly, it aims at prediction, not so much at hypotheses testing. Secondly, it only considers size-independent relationships, where account item values are distributed nearly lognormally in cross-section. This is generally verified with

data extracted from Compustat and other files, which exclude small, unlisted companies. However, in more general cases, distributions may be three-parametric lognormal (Aitchison and Brown, 1957) or even momentless (Ashton *et al.*, 2004), in which case relationships are not loglinear. Therefore, the methodology presented here is at an initial step, and the notional approximation from which it is constructed must be refined and adapted for broader purposes.

While the multiplicative nature of ratio components allows the removal of the effect of firm size when the purpose is financial analysis, it also renders ratios unsuitable when the purpose is the predictive modelling of relationships. The poor effectiveness of accounting-based models stems from a lack of awareness of this contradiction, which also leads to the uncritical use of modelling practices that cannot deliver performance by their very nature. The theoretical developments reported in the paper provide a path forward to overcome this contradiction. They arise from explicitly considering the effect of firm size. When size is treated as a statistical effect, it is straightforward to demonstrate why ratios eliminate it from cross-sections of financial statement data. Similarly, it is easy to demonstrate how a proper modelling methodology can eliminate size while uncovering the set of optimal predictors.

Lev and Gu's (2016) claim that accounting information is losing its relevance must be considered in light of the false limitations that render it infeasible from the outset. The current discredit of accounting-based modelling stems from cognitive biases that have repeatedly been accepted under the guise of scientific rigour. The most damaging examples of this are the dogmatic and erroneous rejection of loglinear modelling (Eisenbeis, 1977), and the unwarranted stigma surrounding sampling designs (Ohlson, 1980; Palepu, 1986), without which accounting information cannot be statistically processed or modelled.

## **Reference list**

Aitchison, J. and J. Brown (1957). *The Lognormal Distribution* (Cambridge Univer. Press).  
Ak, B., Dechow, P., Sun, E. and A. Wang (2013). "The use of financial ratio models to help

- investors predict and interpret significant corporate events", *Australian Journal of Management*, Vol. 38 No. 3 pp. 553-598.
- Altman, E. (1968). "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", *The Journal of Finance*, Vol. 23 No. 4 pp. 589-609.
- Altman, E. (1983). *Corporate Financial Distress: a Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy* (Wiley, New York).
- Altman, E. and G. Sabato (2007). "Modelling credit risk for SMEs: evidence from the U. S. market", *Abacus*, Vol. 43 No. 3, pp. 332–357.
- Amania, F. and A. Fadlallab (2017). "Data mining applications in accounting: a review of the literature and organizing framework", *International Journal of Accounting Information Systems*, Vol. 24, pp. 32–58.
- Ashton, D., Dunmore, P. and M. Tippet (2004). "Double entry bookkeeping and the distributions of a firm's financial ratios". *Journal of Business Finance and Accounting*, Vol. 31, No. 5-6, pp. 583-606.
- Balcaen, S. and H. Ooghe (2006), "35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems", *The British Accounting Review*, Vol. 38, No. 1, pp. 63–93.
- Barnes, P. (1990). "The prediction of takeover targets in the UK with multiple discriminant analysis", *Journal of Business Finance and Accounting*, Vol. 17, No. 1, pp. 73-84.
- Beneish, M. (1999). "The detection of earnings manipulation", *Financial Analysts Journal*, Vol. 55, No. 5, pp. 24–36.
- Bird, R., Gerlach, R. and A. Hall (2001). "The prediction of earnings movements using accounting data: an update and extension of Ou and Penman", *Journal of Asset Management*, Vol. 2, No. 2, pp. 180–195.
- Casey, R., Gao, F., Kirschenheiter, M., Li, S. and S. Pandit (2016). "Do Compustat financial

- statement data articulate?”, *Journal of Financial Reporting*, Vol. 1, No. 1, pp. 37–59.
- Clout, V., Falta, M. and R. Willett (2016). “Earnings in valuation and their value relevance”, *Journal of Contemporary Accounting and Economics*, Vol. 12, No. 3, pp. 223–240.
- Cosslett, S., Manski, C. and D. McFadden (1981). “Efficient estimation of discrete choice models”, in *Structural Analysis of Discrete Data with Econometric Applications*, Vol. 3 pp. 51-111 (The MIT Press, Cambridge, MA).
- Dechow, P., Weili, G., Larson, C. and R. Sloan (2011). “Predicting material accounting misstatements”, *Contemporary Accounting Research*, Vol. 28, No. 1, pp. 17–82.
- Dambolena, I., Eriksen, S., and D. Kopcsó (2009). “Logarithmic transformations in regression: do you transform back correctly?” *PRIMUS*, Vol. 19, No. 3, pp. 280-295.
- Du Jardin, P. (2009). “Bankruptcy prediction models: how to choose the most relevant variables?”, *Bankers, Markets & Investors*, Vol. 98, pp. 39–46.
- Eisenbeis, R. (1977). “Pitfalls in the application of discriminant analysis in business, finance, and economics”, *The journal of finance*, Vol. 32, No. 3, pp. 875-900.
- Eilifsen, A. and W. Messier (2000). “The incidence and detection of misstatement: a review of and integration of research”, *Journal of Accounting Literature*, Vol. 19 pp. 1-43.
- Falta, M. and R. Willett (2011). “Multiplicative regression models of the relationship between accounting numbers and market value”, *SSRN paper no. 2130328*.
- Holthausen, R. and D. Larcker (1992). “The prediction of stock returns using financial statement information”, *Journal of Accounting and Economics*, Vol. 15, No. 2 pp. 373-411.
- Ijiri, I., and H. Simon (1964). “Business firm growth and size”, *The American Economic Review*, Vol. 54 No. 2 pp. 77–89.
- John, J., and N. Draper (1980). “An alternative family of transformations”, *Journal of the Royal Statistical Society, Series C*, Vol. 29, No. 2, pp. 190–197.
- Johnson N., Kotz, S. and N. Balakrishnan (1994). “Lognormal distributions”, in *Continuous*

- Univariate Distributions*, Vol. 1 (Wiley, New York).
- Kotz, S., Balakrishnan, N. and N. Johnson (2005). *Continuous Multivariate Distributions: Models and Applications*, Vol. 1 (Wiley, New York).
- Lau, H. Lau, A. and D. Gribbin (1995). “On modelling cross-sectional distributions of financial ratios”, *Journal of Business Finance and Accounting*, Vol. 22, No. 4, pp 521–549.
- Leone, A., Minutti-Meza, M. and M. Wasley (2017). “Influential observations and inference in accounting research”, *Simon Business School Working Paper FR 14-06*.
- Lev B., and F. Gu (2016). *The End of Accounting and the Path Forward for Investors and Managers* (Wiley, New York).
- Liou, F. (2008). “Fraudulent financial reporting detection and business failure prediction models, a comparison”, *Managerial Auditing Journal* Vol. 23, No. 7, pp. 650-662.
- Manski, C. and D. MacFadden (1981). *Structural analysis of discrete data with econometric applications*. (The MIT Press, Cambridge, MA).
- McLeay, S. and D. Trigueiros (2002). “Proportionate growth and the theoretical foundations of financial ratios.” *Abacus*, Vol. 38, No. 3, pp. 297–316.
- Ohlson, J. (1980). “Financial ratios and the probabilistic prediction of bankruptcy”, *Journal of Accounting Research*, Vol. 18, No. 1, pp. 512–534.
- Ohlson, J., and S. Kim (2015). “Linear valuation without OLS: the Theil-Sen estimation approach”, *Review of Accounting Studies*, Vol. 20, No. 1, pp. 395-435.
- Ou, J., and S. Penman (1989). “Financial statement analysis and the prediction of stock returns”, *Journal of Accounting and Economics*, Vol. 11, No. 4, pp. 295-329.
- Ou, J. (1990). “The Information content of non-earnings accounting numbers as earnings predictors”, *Journal of Accounting Research*, Vol. 28, No. 1, pp. 144–163.
- Palepu, K. (1986). “Predicting takeover targets, a methodological and empirical analysis”, *Journal of Accounting and Economics*, Vol. 8, No. 1, pp. 3-35.

- Singh, A. and G. Whittington (1968). *Growth, Profitability and Valuation* (Cambridge University Press).
- So, J. (1987). “Some empirical evidence on the outliers and the non-Normal distribution of finance ratios”, *Journal of Business Finance and Accounting*, Vol. 14, No. 4, pp. 483–496.
- Sun, J., Li, H., Huang, Q. and K. He (2014). “Predicting financial distress and corporate failure: a review from the state-of-the-art definitions, modelling, sampling, and featuring approaches”, *Knowledge-Based Systems*, Vol. 57, pp. 41-56.
- Taffler, R. (1982). “Forecasting company failure in the UK using discriminant analysis and finance ratio data”, *Journal of the Royal Statistical Society*, Vol. A-145, No. 3, pp. 342-58.
- Tian, S., Yu, Y. and H. Guo (2015). “Variable selection and corporate bankruptcy forecasts”, *Journal of Banking and Finance*, Vol. 52-C, pp. 89-100.
- Tippett, M. (1990). “An induced theory of financial ratios”, *Accounting and Business Research*, Vol. 21, No. 81, pp. 77-85.
- Trigueiros, D. (1995). “Accounting identities and the distribution of ratios”, *British Accounting Review*, Vol. 27, No. 2, pp. 109–126.
- Watson, C. (1990). “Multivariate distributional properties, outliers, and transformation of financial ratios”, *The Accounting Review*, Vol. 65, No. 3, pp. 682-695.
- Wilcox, R. (2010). *Fundamentals of Modern Statistical Methods* (Springer, New York).
- Xu, W., Xiao Z., Dang X., Yang D. and X. Yang (2014). “Financial ratio selection for business failure prediction using Soft Set”, *Knowledge-Based Systems*, Vol. 63, pp. 59-67.
- Xie, Y. and C. Manski (1989). “The logit model and response-based samples”, *Sociological Methods and Research*, Vol. 17, No. 3, pp. 283–302.
- Zmijewski, M. (1984). “Methodological issues related to the estimation of financial distress prediction models”, *Journal of Accounting Research*, Vol. 22, pp. 59-82.

**Table 1**  
**Set of accounts' items used in the three experiments**

|       |                                           |       |                                                        |
|-------|-------------------------------------------|-------|--------------------------------------------------------|
| CHE   | Cash and Short-Term Investments *         | RE    | Retained Earnings *                                    |
| RECT  | Receivables (total) *                     | SEQ   | Shareholders' Equity (total)                           |
| INVT  | Inventories (total)                       | REVT  | Revenue (total) *                                      |
| ACT   | Current Assets (total) *                  | XOPR  | Operating Expense (total) *                            |
| WCAP  | Working Capital *                         | COGS  | Cost of Goods Sold                                     |
| PPENT | Property Plant and Equipment (total, net) | GP    | Gross Profit (loss) *                                  |
| INTAN | Intangible Assets (total)                 | XSGA  | Selling, General and Administrative Expenses           |
| AT    | Assets (total) *                          | DP    | Depreciation and Amortization (total)                  |
| AP    | Account Payable/Creditors (trade)         | OIADP | Operating Income after Depreciation *                  |
| XACC  | Accrued Expenses                          | XINT  | Interest and Related Expense                           |
| TXP   | Income Taxes Payable                      | TXT   | Income Tax (total) *                                   |
| LCT   | Current Liabilities (total) *             | IB    | Income Before Extraordinary Items                      |
| DLTT  | Long-Term Debt (total) *                  | NI    | Net Income (loss)                                      |
| LT    | Liabilities (total) *                     | OANCF | Net Cashflow from Operating Activities *               |
| BKVL  | Book Value of Common and Preferred Stock  | CAPX  | Capital Expenditure                                    |
| PSTK  | Preferred Stock                           | EPS   | Earnings per Share (basic, before extraordinary items) |
| CSTK  | Common/Ordinary Stock *                   | DVPS  | Dividends per Share (ex-date, fiscal) *                |

The table lists the 32 items plus 2 pre-existing ratios tested for inclusion in the first stage of the new methodology, where both current values and changes obeying transformation depicted in (8) in relation to prior year are tested, and the components of ratios tested for inclusion in ratio-based models. Asterisks show items selected in the first stage by any one of the three experiments. Item values, abbreviations and names are taken from the Compustat file, except BKVL which is computed from other items in the same file.

**Table 2**  
**Bankruptcy prediction**

|                                                                                         | Ratio model |        | First-stage model |        |
|-----------------------------------------------------------------------------------------|-------------|--------|-------------------|--------|
| Model-building data, number of non-bankrupt cases and percentage of total               | 845         | 50.12% | 845               | 50.12% |
| Model-building data, number of bankruptcy cases and percentage of total                 | 841         | 49.88% | 841               | 49.88% |
| Model-testing data, non-bankrupt, N, and percentage N / (N+P) of total                  | 837         | 49.76% | 837               | 49.76% |
| Model-testing data, bankrupt, P, and percentage P / (N+P) of total                      | 845         | 50.24% | 845               | 50.24% |
| Overall, correctly classified, TP + TN, and percentage (TP+TN) / (N+P) of total         | 1,601       | 95.18% | 1,619             | 96.25% |
| Number of non-bankruptcies correctly classified, TN, and percentage TN / N              | 782         | 93.43% | 810               | 96.77% |
| Number of non-bankruptcies incorrectly classified, FP, and percentage FP / N            | 55          | 6.57%  | 27                | 3.23%  |
| Number of bankruptcies correctly classified, TP, and percentage TP / P                  | 819         | 96.92% | 809               | 95.74% |
| Number of bankruptcies incorrectly classified, FN, and percentage FN / P                | 26          | 3.08%  | 36                | 4.26%  |
| Precision, TP / (TP + FP)                                                               |             | 93.71% |                   | 96.77% |
| Recall (sensitivity), TP / (TP + FN)                                                    |             | 96.92% |                   | 95.73% |
| Predicted as not bankrupt, (TN + FN) / (N+P), and % difference to model-building        | 48.04%      | -2.08% | 50.30%            | 0.18%  |
| Predicted as bankruptcy, (TP + FP) / (N+P), and % difference to model-building          | 51.96%      | 2.08%  | 49.70%            | -0.18% |
| Cox and Snell pseudo-R-Square                                                           | 0.595       |        | 0.682             |        |
| Chi-Square, degrees of freedom and significance                                         | 1,526       | 5 ***  | 1,931             | 5 ***  |
| ROC curve, predicted probability and standard error                                     | 0.979       | 0.002  | 0.991             | 0.001  |
| X1, ratio of Working Capital to Assets (total)                                          | -1.309      | ***    |                   |        |
| X2, ratio of Retained Earnings to Assets (total)                                        | -0.028      | *      |                   |        |
| X3, ratio of Operating Income after Depreciation to Assets (total)                      | -4.343      | ***    |                   |        |
| X4, ratio of Book Value of Common and Preferred Stock to Liabilities (total)            | -1.015      | ***    |                   |        |
| X5, ratio of Revenue to Assets (total)                                                  | 0.023       |        |                   |        |
| Logarithm of absolute value of Working Capital, minus 7 for negative values             |             |        | 0.075             | *      |
| Logarithm of Assets (total)                                                             |             |        | -9.55             | ***    |
| Logarithm of Liabilities (total)                                                        |             |        | 9.66              | ***    |
| Logarithm of absolute value of Retained Earnings, minus 7 for negative values           |             |        | -0.402            | **     |
| Logarithm of absolute value of Operating Income after Depreciation, minus 7 for neg. v. |             |        | -0.401            | ***    |
| Constant term                                                                           | 1.934       | ***    | 6.539             | ***    |

Top to bottom: sample size, accuracy, predicted class frequencies and differences in relation to model-building class frequencies, goodness-of-fit statistics, and model coefficients. All line items are descaled by  $10^6$  before transformed. 'Logarithm' refers to base 10 logarithm. All zero values are kept as zero. 'T' means correctly classified, 'F' means incorrectly classified. Significances are shown as asterisks: '\*\*\*' is  $p \leq 0.001$ , '\*\*' is  $p \leq 0.01$  and '\*' is  $p \leq 0.05$ .

**Table 3**  
**Financial misstatement detection**

|                                                                                         | Ratio model |        | First-stage model |        |
|-----------------------------------------------------------------------------------------|-------------|--------|-------------------|--------|
| Model-building data (1976-1999), number of not misstated and percentage of total        | 376         | 48.20% | 376               | 48.20% |
| Model-building data (1976-1999), number of misstated and percentage of total            | 405         | 51.80% | 405               | 51.80% |
| Model-testing data (2000-2008), not misstated, N, and percentage N / (N+P) of total     | 312         | 43.57% | 312               | 43.57% |
| Model-testing data (2000-2008), misstated, P, and percentage P / (N+P) of total         | 404         | 56.42% | 404               | 56.42% |
| Overall, correctly classified, TP + TN, and percentage (TP+TN) / (N+P) of total         | 530         | 74.02% | 602               | 84.08% |
| Number of not misstated correctly classified, TN, and percentage TN / N                 | 202         | 64.74% | 272               | 87.18% |
| Number of not misstated incorrectly classified, FP, and percentage FP / N               | 110         | 35.26% | 40                | 12.82% |
| Number of misstated cases correctly classified, TP, and percentage TP / P               | 328         | 81.19% | 330               | 81.68% |
| Number of misstated cases incorrectly classified, FN, and percentage FN / P             | 76          | 18.81% | 74                | 18.32% |
| Precision, TP / (TP + FP)                                                               |             | 74.89% |                   | 89.19% |
| Recall (sensitivity), TP / (TP + FN)                                                    |             | 81.19% |                   | 81.68% |
| Predicted as not misstated, (TN + FN) / (N+P), and % difference to model-building       | 38.83%      | -9.37% | 48.32%            | 0.12%  |
| Predicted as misstated, (TP + FP) / (N+P), and % difference to model-building           | 61.17%      | 9.37%  | 51.68%            | -0.12% |
| Cox and Snell pseudo-R-Square                                                           | 0.335       |        | 0.596             |        |
| Chi-Square, degrees of freedom and significance                                         | 312         | 12 *** | 707               | 9 ***  |
| ROC curve, predicted probability and standard error                                     | 0.821       | 0.011  | 0.947             | 0.005  |
| DSRI, Days Sales in Receivables index, see caption                                      | 0.178       | *      |                   |        |
| SGI, Sales growth = Revenue / Revenue prior year                                        | 0.435       | *      |                   |        |
| DEPI, Depreciation to Property, Plant and Equipment, see caption                        | 0.921       | **     |                   |        |
| SGAI, Sales, General and Administrative Expenses index, see caption                     | 0.439       | *      |                   |        |
| LVGI, Leverage index, see caption                                                       | 0.737       | **     |                   |        |
| Ratio of Net Income to Revenue                                                          | 0.120       | **     |                   |        |
| Receivables growth = Receivables / Receivables in prior year                            | -0.062      | *      |                   |        |
| Current Ratio = Current Assets (total) / Current liabilities (total)                    | 0.142       | *      |                   |        |
| Ratio of Revenue to Assets (total)                                                      | 1.140       | ***    |                   |        |
| Ratio of Cash and Short-Term Investments to Assets (total)                              | -2.298      | ***    |                   |        |
| Ratio of Inventories to Assets (total)                                                  | 1.574       | *      |                   |        |
| Ratio of Liabilities (total) to Assets (total)                                          | -4.426      | ***    |                   |        |
| Logarithm of Cash and Short-Term Investments                                            |             |        | -1.733            | ***    |
| Logarithm of Cash and Short-Term Investments, change $\Delta$ in relation to prior year |             |        | 0.481             | *      |
| Logarithm of Receivables (total)                                                        |             |        | 0.549             | **     |
| Logarithm of Assets (total)                                                             |             |        | 11.31             | ***    |
| Logarithm of Long-Term Debt                                                             |             |        | 0.230             | ***    |
| Logarithm of Liabilities (total)                                                        |             |        | -11.82            | ***    |
| Logarithm of Liabilities (total), change $\Delta$ in relation to prior year             |             |        | 3.760             | ***    |
| Logarithm of Common/Ordinary Stock (Capital)                                            |             |        | -1.80             | ***    |
| Logarithm of Revenue (total)                                                            |             |        | 4.186             | ***    |
| Constant term                                                                           | -1.118      |        | -15.47            | ***    |

Top to bottom: sample size, accuracy, predicted class frequencies and differences in relation to model-building class frequencies, goodness-of-fit statistics, and model coefficients. All line items are descaled by  $10^6$  before transformed. 'Logarithm' refers to base 10 logarithm. All zero values are kept as zero. 'T' means correctly classified, 'F' means incorrectly classified. Significances are shown as asterisks: '\*\*\*' is  $p \leq 0.001$ , '\*\*' is  $p \leq 0.01$  and '\*' is  $p \leq 0.05$ . DSRI=(RECT/REVT) / (RECT/REVT) prior year. DEPI=DEP/(DEP+PPENT) prior year / DEP/(DEP+PPENT). SGAI=(XSGA/REVT) / (XSGA/REVT) prior year. LVGI=((DLTT+LCT)/AT) / ((DLTT+LCT)/AT) prior year.

**Table 4**  
**Earnings' direction forecast**

|                                                                                         | Ratio model |         | First-stage model |        |
|-----------------------------------------------------------------------------------------|-------------|---------|-------------------|--------|
| Model-building data, number of non-increases and percentage of total                    | 41,875      | 64.66%  | 41,875            | 64.66% |
| Model-building data, number of increases and percentage of total                        | 23,303      | 35.34%  | 23,303            | 35.34% |
| Model-testing data, non-increases, N, and percentage N / (N+P) of total                 | 41,762      | 64.66%  | 41,796            | 64.66% |
| Model-testing data, increases, P, and percentage P / (N+P) of total                     | 22,826      | 35.34%  | 22,841            | 35.34% |
| Overall, correctly classified, TP + TN, and percentage (TP+TN) / (N+P) of total         | 48,008      | 74.33%  | 52,006            | 80.46% |
| Number of not increases correctly classified, TN, and percentage TN / N                 | 37,266      | 89.23%  | 35,828            | 85.72% |
| Number of not increases incorrectly classified, FP, and percentage FP / N               | 4,496       | 10.77%  | 5,968             | 14.28% |
| Number of increases correctly classified, TP, and percentage TP / P                     | 10,742      | 47.06%  | 16,178            | 70.83% |
| Number of increases incorrectly classified, FN, and percentage FN / P                   | 12,084      | 52.94%  | 6,663             | 29.17% |
| Precision, TP / (TP + FP)                                                               |             | 70.49%  |                   | 73.05% |
| Recall (sensitivity), TP / (TP + FN)                                                    |             | 47.06%  |                   | 70.83% |
| Predicted as a not increase, (TN + FN) / (N+P), and % difference to model-building      | 76.41%      | 12.16%  | 65.74%            | 1.49%  |
| Predicted as an increase, (TP + FP) / (N+P), and % difference to model-building         | 23.59%      | -12.16% | 34.26%            | -1.49% |
| Cox and Snell pseudo-R-Square                                                           | 0.178       |         | 0.342             |        |
| Chi-Square, degrees of freedom and significance                                         | 12,773      | 10 ***  | 27,297            | 10 *** |
| ROC curve, predicted probability and standard error                                     | 0.753       | 0.002   | 0.86              | 0.002  |
| Ratio of Cash flow from Operating Activities to Liabilities (total)                     | -0.091      | **      |                   |        |
| Ratio of Income Tax to Liabilities (total)                                              | -3.449      | ***     |                   |        |
| Ratio of Retained Earnings to Liabilities (total)                                       | -0.023      | ***     |                   |        |
| Ratio of Net Income to Assets (total)                                                   | -0.063      | **      |                   |        |
| Ratio of Liabilities (total) to Assets (total)                                          | 0.055       | **      |                   |        |
| Ratio of Gross Profit to Assets (total)                                                 | -0.201      | ***     |                   |        |
| Growth in the Revenue to Assets (total) ratio (Ou, 1990)                                | 0.030       | **      |                   |        |
| ROR, Income Before Extraordinary Items to Shareholders Equity prior year (Ou, 1990)     | -0.008      | **      |                   |        |
| ROR in prior year (Ou, 1990)                                                            | 0.005       | *       |                   |        |
| Logarithm of Assets (total)                                                             | -0.685      | ***     | -0.614            | ***    |
| Logarithm of Liabilities (total)                                                        |             |         | 0.643             | ***    |
| Log-Modulus of Retained Earnings                                                        |             |         | -0.106            | ***    |
| Log-Modulus of Retained Earnings, change $\Delta$ in relation to prior year             |             |         | 0.036             | ***    |
| Log-Modulus of Gross Profit                                                             |             |         | -0.094            | ***    |
| Log-Modulus of Gross Profit, change $\Delta$ in relation to prior year                  |             |         | 0.053             | ***    |
| Log-Modulus of Income Tax                                                               |             |         | -0.060            | ***    |
| Logarithm of 1,000 times Dividends per Share ex-date, fiscal year-end                   |             |         | -0.208            | ***    |
| Logarithm of 1,000 times Dividends per Share, change $\Delta$ in relation to prior year |             |         | 0.110             | ***    |
| Log-Modulus of Cashflow from Operating Activities                                       |             |         | -0.043            | ***    |
| Constant term                                                                           | 5.080       | ***     | 0.702             | ***    |

Top to bottom: sample size, accuracy, predicted class frequencies and differences in relation to model-building class frequencies, goodness-of-fit statistics, and model coefficients. All line items are descaled by  $10^6$  before transformed. 'Logarithm' refers to base 10 logarithm. All zero values are kept as zero. 'T' means correctly classified, 'F' means incorrectly classified. Significances are shown as asterisks: '\*\*\*' is  $p \leq 0.001$ , '\*\*' is  $p \leq 0.01$  and '\*' is  $p \leq 0.05$ . The Log-Modulus transformation of  $x$  is  $\text{sign } x \log_{10} |x|$ .

**Figure 1**  
**ROC curves**

Figure 1A, bankruptcy prediction.

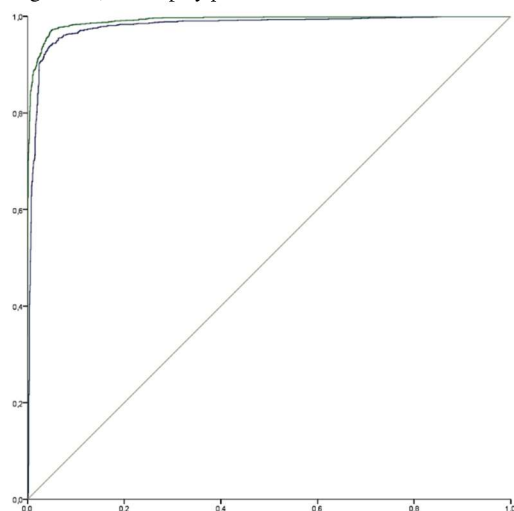


Figure 1B, misstatement detection.

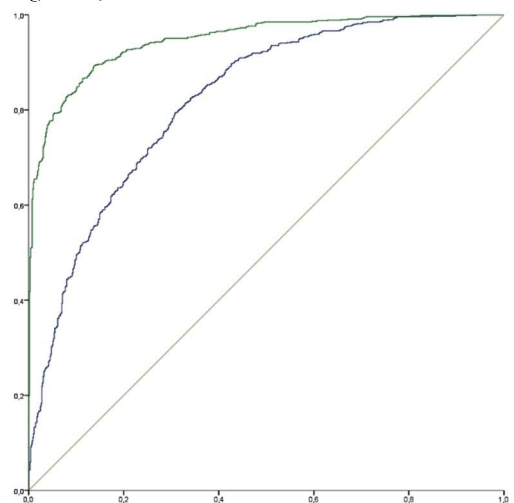
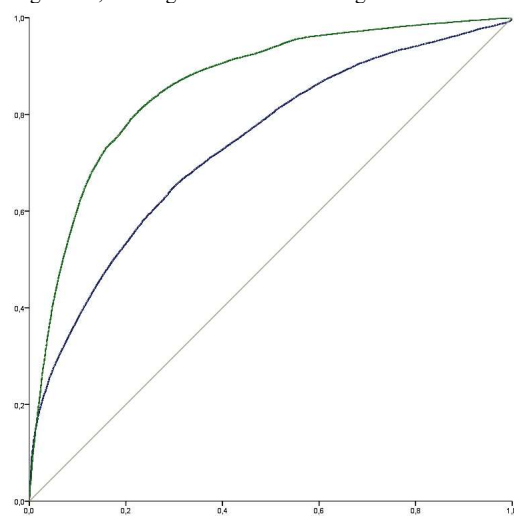


Figure 1C, Earnings direction forecasting.



The ROC curves compare ratio (blue) and second stage (green) models. The X-axis is the specificity of the positive class (bankruptcy, misstatement, earnings) and the Y-axis is the sensitivity (bankruptcy, misstatement, earnings).