The use of accounting data in statistical models

Duarte Trigueiros

Escola Superior de Economia - Universidade do Algarve

Resumo

É difícil usar dados contabilísticos em modelos estatísticos pois abundam os casos influentes e as distribuições são heteroscedásticas, leptocúrticas e fortemente assimétricas. O objectivo deste artigo é fornecer sugestões que facilitem a compreensão do comportamento estatístico destes dados de modo a permitir uma mais fácil modelação e teste de hipóteses.

Palavras Chave: Dados contabilísticos, modelação estatística

Abstract

It is difficult to use accounting data in statistical models. Outliers abound, influential cases distort estimation while heteroscedasticity, fat tails and asymmetric distributions make P-values meaningless. When trying to solve these problems researchers find out that accounting data is not only difficult to use, its statistical behaviour is also difficult to predict.

The goal of this article is to facilitate understanding of the statistical characteristics of accounting numbers. Guidelines are provided on how to transform raw data into well-behaved variables, how to deal with non-proportionalities, how to accurately model firm size and solve other difficulties.

Keywords: Accounting data, statistical modelling

1 Accounting Numbers are Multiplicative

Accounting data is the numerical information found in annual reports of firms. Reports contain sets of accounts such as the Profit and Loss Account, the Balance sheet and others. Each item in these sets reports on a specific magnitude. The volume of sales of the year, for instance, is reported in the item named Sales. Magnitudes found in sets of accounts are the raw material for financial analysis and all types of statistical manipulations. But before being of any use these numbers are usually combined to form ratios. Typically, two numbers from the same report say Earnings and Net Worth may be chosen as the most relevant and determination of a ratio. Most accounting data used in statistical models is in the form of ratios.

It is impossible to understand the statistical characteristics of ratios without understanding first the characteristics of the numbers they are made of and the way they interact. Therefore, the following lines are devoted to such numbers.

The first and most important fact about numbers found in accounts is that they cannot be described as resulting from the type of random mechanism that leads to Normal variables. Normal variables stem from additive random mechanisms. An observed distribution is the result of adding a large number of other distributions. For instance, the distribution of weight in adults stems from adding the probabilities associated with genetic effects, eating habits and other effects. In the limit, any addition of probability distributions, no matter which, leads to the Normal distribution. This is known as the Central Limit theorem.

Obviously, in an additive mechanism, each intervening effect may lead to an increase or to a decrease in the likelihood associated with the resulting event. Adequate eating habits may, for instance, be able to balance a genetic predisposition to put on weight, lowering the expected weight. The mechanism generating accounting numbers is different: effects always reinforce each other. This is because such effects are, in this case, every individual transaction contributing to a reported magnitude. Indeed, each transaction contributing to the amount reported as, say, Sales for a given period, is itself a random event. It contributes to the reported number, not in a manner, which could lead to either an increase or decrease in Sales, but by accumulation only.

Accumulations of random events lead to multiplicative, as opposed to additive variables, because the likelihood of realisations is conditional on the occurrence of a chain of previous events, not on any free interaction of influences, some positive, others negative. Such likelihood thus stems from multiplying rather than adding probabilities.

Multiplicative distributions are easy to recognise. They are skewed, exhibiting long tails towards positive values. As a consequence, some of the observations in a sample are likely to exhibit very large magnitudes in comparison with others, thus giving the impression of being outliers.

Contrasting with additive distributions where skewness and kurtosis are independent, in mechanisms of the Multiplicative type both statistics are manifestations of a unique, underlying phenomenon, variability. Therefore, highly volatile variables exhibit markedly skewed and leptokurtic distributions whereas those where variability is small have almost symmetrical, non kurtotic distributions.

Accumulation is just one amongst several processes leading to Multiplicative numbers. Any variable where magnitude is, on average, proportional to changes dx, will exhibit a Multiplicative distribution. The natural form of the origin of such type of variable is

$$\int \frac{dx}{x} = \ln x + c \qquad (1)$$

where μ is an expected percent change, x is the variable supposed to drive changes in x, dε is a small random disturbance and μdt is supposed to be independent of x. Therefore, in this type of variable, percent changes are additive. Normality, as a term, is approached by percent change, not by absolute change.

The mechanism depicted in (3) is known as the "Gibrat's Law or Proportionate Effect". It leads to lognormal distributions (distributions where the logarithm of observations is Normal) or to other types of multiplicative distribution. Multiplicative proportionate exponential and lognormal are terms variously used to designate the family of skewed distributions with its origin in (1). Aitchison and Brown (1957) describe the lognormal distribution.

Sales, Earnings, Assets and other accounting aggregates are since long known, in domains such as Industrial Economics and others, to be multiplicative, i.e., broadly lognormal. In spite of this, until recently the accounting literature has discarded lognormality in accounting data as incompatible with sound reasoning while quoting the influential fallacy introduced by Eisenbeis (1977) or the fact that some ratios are apparently Normal, or even the existence of negative values in some accounts. Excessive skewness and other characteristics of multiplicative variables were interpreted as distortions of normality or, in the case of ratios, as a side effect of non-proportionality (Barnes, 1982). In the Accounting research domain, the case for multiplicative mechanisms was made by McCrae (1986) and by Trigueiros (1995).

The peculiar characteristics of lognormal variables must be borne in mind in any context involving the manipulation of these variables or their ratios. Lognormality cannot be treated as a simple departure from normality. For coefficients of variation beyond 0.25, skewness and kurtosis are so severe that most observations concentrate in a small region with only a few extreme values spreading out over a wide range. No parametric tool is robust enough to avoid severe distortion when such data is used.

Statistical models are functional forms supposed to reflect inter-relationships amongst effects. Descriptions of inter-relationships amongst effects greatly differ between additive and multiplicative variables. For additive data, distributions are preserved when variables are added or subtracted. This is not the case for multiplicative data where distributions are preserved when variables are multiplied or divided. The addition or subtraction of two Normal variables will be Normal; the product or ratio of two lognormal variables will be lognormal. The simplest additive formulation would be $x = \mu + \varepsilon$ where x is explained as effect μ, the expectation, plus a random deviation ε. The multiplicative equivalent would be $x = \mu\varepsilon$, where x is now explained as the product of a constant magnitude, μ, by a random factor ε.

The likelihood that x may stretch beyond two or three standard deviations above or below μ is very small. Therefore, in general, additive variables describe deviations from an average magnitude but they are unable to describe large differences in magnitude. By contrast, in the case of multiplicative formulations, the exponential nature of ε leads to likely values of x over a much wider range. The volume of sales of United Biscuits, a firm in the 95th size percentile of its industry, is about 500 times that of firms in the 5th size

percentile. Additive observations would never be able to model such huge discrepancies. This is why logarithmic rather denotes size influence whereas normality generally denotes a size-free variable.

2. How to Use Financial Ratios in Statistical Models

The existence of a size influence in magnitudes reported in the accounts of firms led to the use of ratios. Ratios such as Return on Equity, Interest Cover, Debt to Net Worth and many others are widely used by managers, practitioners and analysts. They control for size so that comparisons may be made.

Where accounting numbers are lognormal, then ratios should be lognormal as well. But some ratios show unexpected characteristics, which makes them difficult to use in statistical models. For instance, although most ratios are indeed lognormal, Total Debt to Total Assets, Net Worth to Total Assets and others are apparently Normal. Current Assets to Total Assets is negatively skewed (see, e.g., So, 1987). How is this possible? The reason is straightforward. Accounting identities preclude some ratios from taking on values above or below a certain value (usually 0 or 1). Ratios whose values are skewed or truncated are usually those whose numerator is a component of the denominator. For example, Current Assets is a component of Total Assets, and Net Worth is a component of Total Liabilities plus Net Worth. Ratios of this form are bounded by 0 and 1 if the numerator is smaller than the denominator, and by 0 and infinity if the numerator is larger than the denominator. Ratios of this form are bounded by 0 and 1 if the numerator is smaller than the denominator, and by 0 and infinity if the numerator is larger than the denominator.

Figure 1 Bivariate distribution of Total Assets (X-axis) and Net Worth (Y-scale)

Adequate countermeasures can take into account constraining mechanisms visibly or unsuspected ratios. For example, a ratio where the numerator cannot be larger than the denominator (i.e., ratios of the form

$$\frac{1}{\sum_{i}}$$

can be transformed into the corresponding unconstrained ratio

$$\sum^{-1}_{i}$$

The unconstrained ratio corresponding to the ratio Fixed Assets to Total Assets (FA/TA) is the ratio TA/FA where CA=TA-FA. The information contained in both ratios is the same. The difference between them is just functional.

Table 1 summarises transformations able to bring ratios affected by several types of constraints into parametric behaviour. Ratios where there is no constraint are ignorant.

Table 1: Transformations adequate to be used in constrained ratios (from Mc Leay and Trigueiros, 2003).

Case No	Ratio	Example	Transformation to use	Boundaries of the ratio
1	R (Y/X)	Current Ratio	Log R	]0;∞[
2	R (X-Y)/X	Sales Margin	Log (1-R)	]-∞;1[
3	R (Y-X)/X	Change in Capital Employed	Log (1+R)	]-1;∞[
4	R (Y-X)/X	Interest Cover	Log (R-1)	]1;∞[
5	R X/(Y-X)	Liabilities Ratio	Log (1/R+1)	]0;1[
6	R X/(Y-X)	Leverage Ratio	Log (1/R+1)	]0;∞[

Awareness of the existence of constraining mechanisms and the way they affect ratios removes one major obstacle in understanding and using financial ratios. But not all is explained. A fact that remains unaccounted for is the existence of fat tails (leptokurtosis) in the logarithms of all types of ratios that is, even after appropriate transformations are applied. Such leptokurtosis, though, is not severe and may, in most cases, be ignored.

Before moving on to the following section it is important to recall two facts about ratios. First, the use of ratios requires caution in order to avoid other, well-documented limitations. Ratios, for instance, produce ambiguous results when denominators, not only numerators, may take on negative values, ratios are sensitive to atypical magnitudes (numbers from the Profit and Loss account must be corrected for reported periods different from one year, accounts from the Balance Sheet may be distorted by seasonal effects). The second important fact is that, notwithstanding the above limitations, ratios do not deserve the negative image conveyed by scientific journals and books. In spite of their widespread use, ratios are regarded with scepticism by scholars and are described as some primitive tradition with no scientific support. One major cause of such scepticism is the unexpected diversity in ratios' empirical distributions. Another is the fact that some reported

numbers may take on negative values, which is difficult to reconcile with multiplicative behaviour.

The following section discusses the distribution of accounting flows or differences (such as Earnings) which may indeed take on negative values.

3. The Distribution of Earnings

The literature on the distribution of ratios seems to consider profits as additive, albeit not necessarily Normal (McLeay, 1986, Tippett, 1990, amongst others). Probably this is because authors focus on profitability ratios, not on profits.

Profitability ratios basically express percent changes in Net Worth. Authors believe that a ratio of the form a/b should approach normality since lognormal e lead to Normal da (where da are small disturbances). However, in order to qualify as a small disturbance, da must be really small when compared with a. Now, for most numbers taken from sets of accounts this is simply not the case. Flows such as changes in Net Worth are not small enough to qualify as a disturbance, not least when compared with the respective stock. Annual Sales, in spite of being a flow, often is larger than stocks such as Fixed Assets, e.g., in high turnover industries. Actually, accounting flows are not that different from stocks, being large accumulations. Thus profitability and other ratios are indeed multiplicative.

Numbers that may take on negative values are simply the result of subtracting positive-only accumulations. A magnitude reported as Earnings is obtained by subtracting the different types of costs and expenses from Revenues. The distribution of these variables should therefore stem from subtracting lognormal distributions. The task of analytically determining such distributions is not easy as it requires working out the logarithm of a subtraction. There is however a fact that simplifies analysis. Costs and Revenues are correlated because both are influenced by the same effect, size. When correlation is taken into account it becomes possible to approximate analytically the distribution of flows. It turns out that, for conditions typically found in industries, such distribution is not unique. Rather, it is a juxtaposition of two approximately lognormal density functions, one for positive and the other for negative values, the latter being a mirror-image of the lognormal as depicted in figure. Simulation confirms this result.

Figure 2: The density function of Earnings is a juxtaposition of two lognormal distributions.

Ratios formed with such distributions may be markedly two-tailed, giving the impression that they are near symmetry. Fat tailed distributions such as Student's *t* or Cauchy's may indeed fit them closely (McKay, 1986). It should be made clear, however, that the hypothesis of additive distributions leads to unacceptable conclusions. The reporting of immaterial values must be less likely than that of size-related values. The probability density of Earnings, for instance, must decrease when approaching zero and then increase again after passing through zero into negative values as predicted by Figure 2. This is because losses, as well as profits, must be proportionate to size. Additive distributions would imply that the reporting of immaterial profits or losses is more likely than that of material losses.

Ratios are multiplicative no matter their components' type, stocks or flows. When the ratio is constrained by some accounting identity, an appropriate transformation will bring it back to a standard behaviour. Specifically, profitability ratios will benefit from transformation 3 in Table 1. Indeed, most commonly found situations involving negative flows are solved simply by using this or other transformation. As for ratios where the denominator, not just the numerator, may take on negative values, it seems as though there is no other choice but to consider two populations, one for positive and the other for negative denominators. This, after all, is how practitioners deal with such ratios.

4. How to Account for Non-Proportionality in Ratios

Measurement using ratios requires proportionality between components. If the natural relationship between ratio components x and y is of the form $y = ax + b$ (non-proportional) rather than $y = ax$ (proportional), then the measurement will be misleading as we cannot have constant standards or norms (see, e.g., Lev and Sunder, 1979). The existence of non-proportional ratios is established by Sudarsanam and Taffler (1995) and by others.

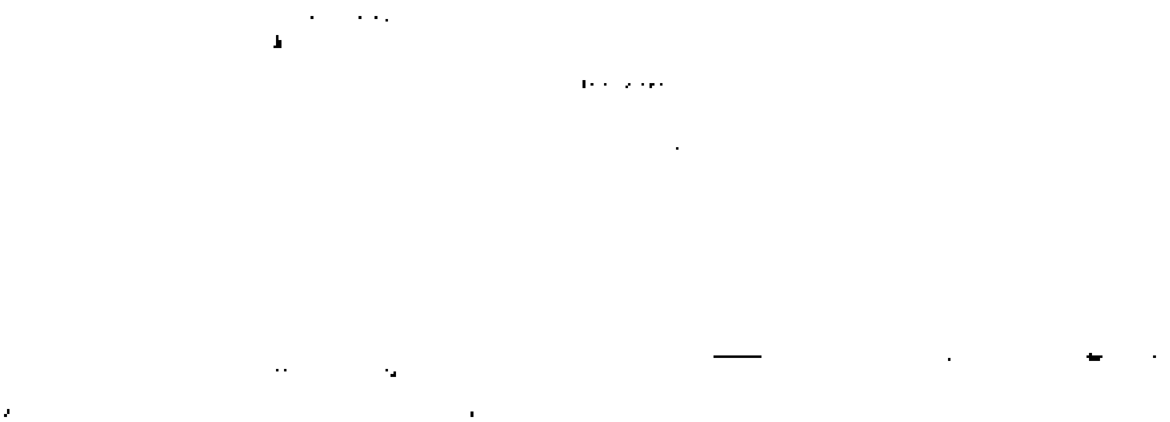
How to overcome problems posed by non-proportionality? The Law of Proportionate Effect acknowledges that changes dx or dy may be proportional, not to x or y, but to $x + \delta$. In this case, the natural form governing the generation of reported magnitudes will be:

$$[illegible] \qquad (2)$$

instead of (1), in them δ may be either constant or size related. Constant δ stems e.g. from the effect of fixed costs in a cost structure. Indeed, in a time series the existence of fixed costs leads to a constant displacement in the distribution of Operating Costs. Size related δ arise in similar cases in cross-section.

Where δ is constant, conditions leading to lognormality [illegible] generate [illegible] distributions known as three-parametric lognormal with threshold δ (Aitchison and Brown, 1957). Three-parametric lognormal density functions are simply the result of displacing lognormal distributions by δ (Figure 3).

Figure 3: The lognormal (solid line) and the three-parametric distributions (dashed lines).



In order to correct [illegible] errors and non-proportionality, one of the following ratios

$$[illegible]$$

should be used for, respectively, three-parametric lognormal numerators or denominators. Thresholds are often δ. Indeed, there exists a [illegible] of such δ for which the above non-proportional ratios meet this proportionality.

How damaging thresholds are δ in the ratio measurement? Due to the exponential character of (2), reported magnitudes may attain values many times larger than the threshold and in such cases [illegible] δ [illegible]. Non-proportionality is significant only where magnitudes are not much larger than δ, for example, in the time series context. Indeed, comparatively large size-independent thresholds are plausible [illegible]. Observations in a time series have their origin in the same object, one firm in different

periods. Values that are constant inside firms, such as fixed costs, may create conceptually large thresholds. In cross-section, as observations have their origin in different objects, size-independent thresholds would require the existence of industry-wide "fixed" costs. Since such costs should allow for the survival of small firms, they must be small. An industry-wide cost of £4m for food manufacturers in the UK would be only 0.2% of United Biscuits' revenues, but it would equal or exceed the turnover of the 5th smallest firms in the industry.

As mentioned, fixed costs are also likely to generate size-related thresholds, mainly in cross section where large firms have large fixed costs and small firms have small fixed costs. In this case, the behaviour of δ is similar to that of any accounting variable where correlation with size is the rule. According to (2), τ will be larger than expected for comparatively large δ, (e.g., the case of large firms) and τ is not constant. As a consequence, size-related thresholds do distort ratio measurement. Notice that the problem, in this case, is not any displacement in the distribution of ratio components. Since δ are small for small c and large for large c, distributions are not displaced. The problem is non-linearity in their relationship. Indeed, size-related thresholds require the use of ratios of the type

$$\frac{c}{c+\delta}$$

For a specific value of β, the ratio will have a constant standard thus allowing measurement. On a logarithmic scale, the functional form of such measurement is

$$\log y = \beta \log x + \rho + \varepsilon \qquad (3)$$

where ρ is the logarithm of the ratio standard and ε is the observed deviation from that standard. (3) is similar to a regression. The slope β is approximate in the unit in the case of strict proportionality. Slopes smaller than 1 mean negative δ. In cross section, they turn large firm's ratios downwards, mimicking scale effects.

How should the two types of δ just outlined (constant and size-related) be estimated? The ratio Fixed Assets (FA) to Current Assets (CA) is now used in a cross section example. Five models of ratio, as follows, are compared. Each model is presented together with its logarithmic counterpart as multiplicative formulations require logarithmic scaling prior to coefficient estimation. Figure 4 shows on a logarithmic scale how the usual ratio method fares compared with each model. Figure 5 reproduces figure 4 on the original scale (only the region near the origin is displayed).

Figure 4: The usual ratio (solid line) compared, on a logarithmic scale, with the slope ratio (Model 2), threshold ratios (Models 3 and 4), and the threshold plus slope ratio (Model 5).

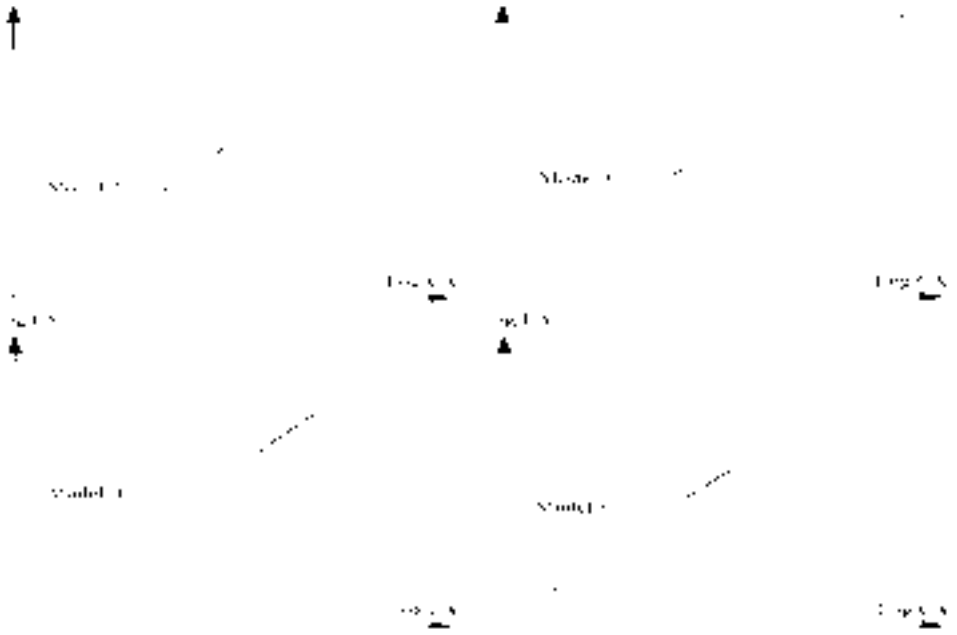
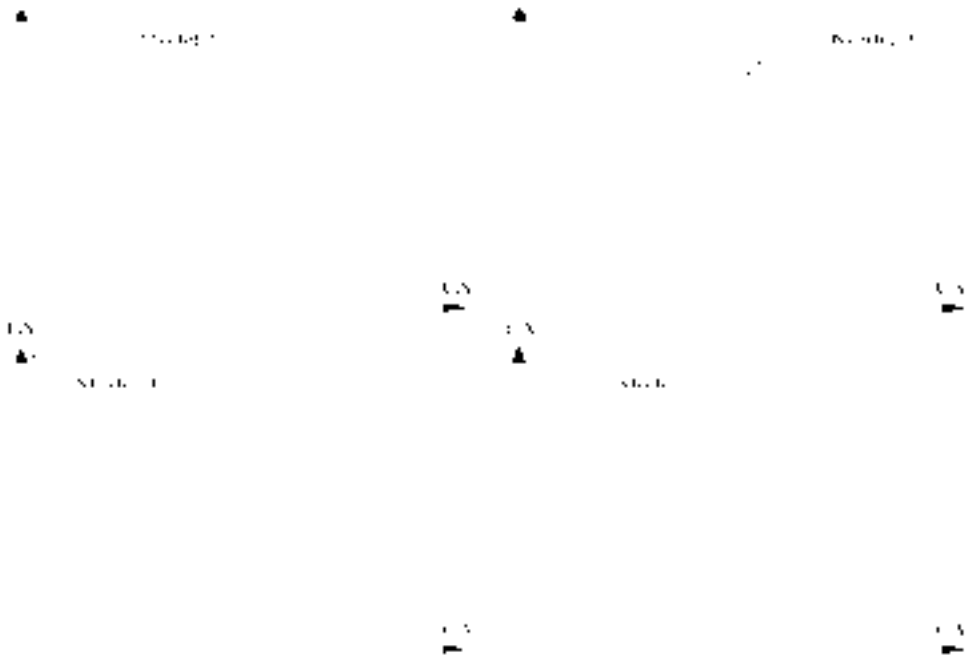


Figure 5: The usual ratio (solid line) compared with the slope ratio (Model 2), threshold ratios (Models 3 and 4), and the threshold plus slope ratio (Model 5).



Model 1: the usual ratio, no correction introduced. This ratio requires the estimation of one parameter, the size band. In cross-section, an appropriate standard is the median of the distribution of the ratio over a reasonable size band. The ratio

$$\frac{FA}{CA} \quad \text{for all } \log(CA) - g < \log(A) < \log(CA) + g$$

with g being a random distribution and $g = 10^{-3}$. When the distributions of CA and FA are nearly exponential, the standard may be estimated by subtracting the averages of log FA and log CA. However, since a significant threshold in the distribution of CA is removed, such a standard will be rarely satisfied.

Model 2: ratio with size-related threshold in the denominator. This ratio may control for the existence of economies of scale and other forms of non-linearity. It requires the estimation of two parameters, the standard up to a constant α and the slope β. Regressions such as (1) or (3) may be used to estimate both.

[illegible]

[illegible] corresponding to the ratio [illegible]

Graphically the ratio [illegible] straight line on logarithmic scale and [illegible]

Model 3: ratio with constant threshold in the denominator (joint estimation). The ratio corrects for a constant threshold [illegible] in such a way that the standard [illegible] as well. It requires the estimation of two parameters: the standard [illegible] and the constant threshold. Both are estimated using:

[illegible] corresponding to the ratio [illegible]

Graphically, the ratio [illegible] on logarithmic scale [illegible] straight line on ordinary scale.

Model 4: ratio with constant threshold in the denominator (independent estimation). In this case there [illegible] degree of freedom [illegible] logarithm of the [illegible] standard. The constant threshold [illegible] the distribution of [illegible] is estimated previous to that of the standard [illegible] using:

[illegible] corresponding to the ratio [illegible]

Graphically [illegible] model 3 [illegible] logarithmic scale and [illegible] ordinary scale. This ratio [illegible] the most [illegible] most important [illegible] sized and large firms.

Model 5: ratio with both constant and size-independent thresholds. This ratio requires the estimation of three parameters [illegible] using:

[illegible] corresponding to [illegible]

Graphically [illegible] models 3 and 4.

The independent estimation of a constant threshold (model 4) may be carried out using any of the procedures to detect three-parameter lognormality in distributions. In this example, $\hat{\delta}$ [illegible] is estimated using the procedure suggested by Royston [illegible]. The [illegible] of the standard (models 3 and [illegible]) is carried out using [illegible]

P. 234

Figure 2 shows the variability explained (R^2) by each of the five ratios. Skewness and kurtosis of residuals on a logarithmic scale (cross-plots) displayed. As can be seen by following plots model 2 (R^2) approaches the variability explained by constant thresholds models (model 1). In Dhrymes & is accounted for, β returns to an estimated value of nearly 1 (model 5).

Table 2: Parameters and statistics of the five models.

Model	α	β	γ	k	Skewness	Kurtosis
1	[illegible]			[illegible]	[illegible]	[illegible]
2	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]
3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Nonetheless, despite the adequacy of the fixed ratio rule, we have five practical time-series regressions where Operating Costs explains Sales and traditionally used to estimate Fixed Costs (the intercept term). Where the thresholds ratios are used as an alternative to such regressions, the obtained estimates are clearly different in magnitude. Where the correlation between Sales and Operating Costs is not very high, then the slope of the regression is clearly smaller than the ratio standard. In the limit, for a correlation approaching zero, such a slope would also become zero and the intercept, which is supposed to estimate all Fixed Costs, would equal the average value of Operating Costs. Regression analysis, therefore, is not the right tool for this task. They introduce in the estimation the spurious effect of correlation. By contrast, the threshold ratio correctly evaluates Fixed Costs as a time-invariant of the distribution of Operating Costs.

5 How to Model Firm Size

Firm size is often assumed as a substitute for numerous theoretical constructs ranging from risk to figures of even political goals. Size is also an ingredient of the many theoretical models. In spite of this widespread use, size has remained a poorly defined concept. Where the use of size is required by theory, empirical studies typically refer to easily available data (Total Assets, Market Capitalization, Sales, etc., e.g. Hoshi and Richardson, 1987).

The counterintuitive character of magnitudes reported in accounts suggests, as stressed ref. to, that the derivation of the distribution is driven by size. It is indeed possible to deduce a simple and effective definition of firm size from the rule by making assumptions. First, reported magnitudes are ratios (Gibrat's Law). Proportionate effects, such as Gibrat's, do not remove the effect of size.

In order for ratios to be effective, the likelihood of observed discrepancies in relation to the standard must be independent of size. For instance, the Return on Assets ratio is reckoned an increase or decrease of 2% has different meanings for small and large firms. The probability distribution of ratios must therefore be homoscedastic in terms of size. This implies, in the general case, that percent changes in both the numerator and denominator must be size independent. As mentioned, variables where percent changes are size independent are said to obey the Gibrat's Law of Proportionate Effect. Indeed, the wide-spread use of ratios agrees with the fact that reported numbers are multiplicative.

It was also mentioned that multiplicative mechanisms lead to broadly lognormal distributions. In fact, observed x generated as in (1) may be described functionally as

$$x = A[1 + \varepsilon]^{t} \quad \text{for discrete } t \text{ or } x = A\exp[\varepsilon t + \varepsilon] \text{ for continuous } t \tag{4}$$

where x is the variable which drives changes in x, ε is a logarithmic expectation and ε is a random increment. The level A is the value of x for $t = 0$. It may be demonstrated that, only where exact ratios compounding is assumed, as in the right-hand side of (4), may ratios be validly used.

As for the second assumption mentioned above, that of ratio y/x eliminating the effect of size, it leads to the requirement that the numerator y and denominator x should both be generated under the effect of equal rates of change ($\varepsilon_y = \varepsilon_x$). In fact, where $\varepsilon_y \neq \varepsilon_x$ the ratio standard would not be constant, showing a rate of change of $\varepsilon_y - \varepsilon_x$ with t. This requirement reputedly suggests that, not just y and x but the other numbers found in a specific annual report are generated under the effect of the same rate of change. Indeed, the validity of the ratio method rests on the validity of several, widely used ratios (not just on one or two cases), where numbers used to form one of such ratios are also used to form other ratios. Therefore, if ratios are to be of any use, there must exist a common source of variability underlying numbers reported in the accounts of a firm in a given year. Where numbers $x_1, x_2, \ldots, x_n$ all belong in a specific annual report, this assumption leads to $\varepsilon_{x_1} = \varepsilon_{x_2} = \ldots = \varepsilon_{x_n}$. It is possible to show that such unique source of variability possesses the attributes of size.

How can size be estimated from this rate of change? A specific annual report, say, report i, is characterised by what value τ assumes, i.e. the magnitude reported by item k is explained as

$$x_{ik} = A_{ik}\exp[\varepsilon_k \tau + \varepsilon]$$

where σ_j is the effect of size (the same for all items reported in j). In an additive form:

$$\log x_{ij} = \mu_k + \sigma_j + \varepsilon \qquad (5)$$

where $\mu_k = \log A_k$ and $\sigma_j = \varepsilon_j$. Formulation (5) is basically an Analysis of Variance, i.e., a type of linear model aimed at explaining variability in terms of membership of discrete classes. Specifically, in (5) $\log x_{ij}$ is explained by its membership of two classes, the item class, μ_k, and the annual report class, σ_j. The item class as a fixed (deterministic) effect, as it denotes the fact that k is a specific item amongst those in the sets of accounts reported by firms. These accounts are indeed fixed in number and in type. By contrast, the annual report's class is a random effect: it denotes the fact that j is one of the (randomly selected) annual reports in the sample. Each of these two classes possesses levels: namely, there can be as many levels of k as items in the sets of accounts; there can be as many levels of j as different reports in the sample.

In cross section, μ_k is the expected value of logarithmic magnitudes reported in item k, estimated as the mean of k calculated using all the annual reports in the sample. Thus, in this case, A_k in (5) is the median of magnitudes reported in k. The size effect, σ_j, is the expected $\log x_{ij} - \mu_k$ for numbers in j and its estimation is straightforward: given N numbers, all of them reported in j, (5) is first applied to each of these numbers. The N formulations obtained are then added. Since σ_j is the same in all of these formulations, it is possible to write

$$\sigma_j = \frac{1}{N}\sum_{i=1}^{N}(\log x_{ij} - \mu_k) + \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i$$

Any source of variability common to all $\log x_{ij}$ in j is, by construction, accounted for by σ_j. Therefore, even where correlation amongst some ε may exist, the term

$$\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i$$

should tend to zero with an increasing N, leading to:

$$\text{Estimate of } \sigma_j = \frac{1}{N}\sum_{i=1}^{N}(\log x_{ij} - \mu_k)$$

An estimated size may thus be obtained simply by averaging the logarithms of appropriately adjusted magnitudes drawn from the firm's actual report. Exact confidence intervals for σ_j can also be obtained, the corresponding standard errors being t-distributed with $N-1$ degrees of freedom.

The estimation of σ_j faces two obvious difficulties. First, x_{ij} cannot be drawn from all possible accounts because items such as Earnings, being a subtraction, may take on negative values and cannot be transformed into logarithms. This pre-selection of items introduces a bias in the estimation of size. Second, inside each annual report, some ε are correlated. This increases the standard error. In practice, however, size can be estimated with accuracy by averaging the logarithms of positive-only magnitudes such as Cash and

Short Term Investments, Receivables, Total Inventories, Property, Plant and Equipment (net), Common Equity, Number of Employees, Net Sales, Cost of Goods Sold, Depreciation, Depreciation of the year, Interest Expense, Dividends

6. Concluding Remarks

Accounting data obeys a set of simple rules. The first rule states that since reported numbers and ratios are subject to error, forecasting models should be used to bound data to defensive behaviour. The second rule explains that accounting identities often distort otherwise multiplicative distributions of ratios into unexpected shapes. Such identities should be accounted for prior to logarithmic transformation. The third rule explains the different ways to account for non-proportionality of errors. The fourth rule states that numbers in a specific report are generated under the influence of size, being possible to obtain such estimate just by averaging several of these numbers conveniently adjusted. The distribution of firm sizes and the determinant of the Jacobian of pools by quintiles were also discussed.

It must be noted now, as many contributions to the literature have, that to a pessimistic view of ratios and accounting data in general. Reasons might be found in their apparent lack of theoretical device, probably fed by an over-concern that accounting rules should be complex and full of exceptions just as the prevalence of such rules is indeed complex and full of exceptions. It is expected that by means of pooled cross-sectional and panel estimators one find that such rules are not as arbitrary as supposed.

References

Aitchison, J. and Brown, J. A. C. (1957), *The Lognormal Distribution*, (Cambridge University Press).

Lawrence, R. J. (1977), 'Pitfalls in the Application of Discriminant Analysis in Business, Finance and Economics', *The Journal of Finance*, Vol. 32, No. 3, pp. 857–870.

Lev, B. and Sunder, S. (1979), 'Methodological Issues in the Use of Financial Ratios', *Journal of Accounting and Economics*, December, pp. 187–210.

McLeay, S. (1986), 'The Role of Means, the Means of Ratios and Other Benchmarks', *Accounting and Business Research*, Vol. 17, No. 1, pp. 79–90.

McLeay, S. and Trigueiros, D. (2002), 'Proportionate Growth and the Theoretical Foundations of Financial Ratios', *Abacus*, Vol. XXXVIII, No. 3, pp. 297–316.

Royston, J. P. (1982), 'An Extension of the Shapiro and Wilk's Test for Normality to Large Samples', *Applied Statistics*, Vol. 31, No. 2, pp. 115–124.

So, J. C. (1987), 'Some Empirical Evidence on Outliers and the Non-Normal Distribution of Financial Ratios', *Journal of Business Finance & Accounting*, Vol. 14, No. 4, pp. 483–496.

Stevenson, P. and Tuttle, R. (1981), 'Financial Ratio Proportionality and Inter-Temporal Stability: An Empirical Analysis', *Journal of Business and Finance*, Vol. XIX, pp. 45–60.

Upton, W. S. (1986), 'A Unified Theory of Financial Ratios', *Accounting and Business Research*, Vol. 16, No. 61, pp. 11–23.

Trigueiros, D. (1984), 'Accounting Identities and the Distribution of Ratios', *British Accounting Review*, Vol. 27, No. 2, pp. 107–129.

Notes

1. Natural forms closer to mechanisms rather than observations. They are often expressed as differentials or as difference equations.

2. Tuschnets (1977) mistakenly stated that 'log-transformed variables give less weight to equal percentage changes in a variable if the raw values are large than when they are smaller. The implication would be that one does not believe that, for example, a $1 million difference between a $1 billion and a $2 billion size firm is the same as between a $1 million and a $2 million size firm. The percentage difference is the same in the latter as in the former case (in $'s). Essentially, what is proposed is that the ratio between proportions. The log transformed measurement is equivalent to calculating proportions before. This is appropriate when working against logarithmic transformations were suggested at the time of the use of raw ratio techniques such as those proposed by Deakin and Horrigan (1968).

5. The coefficient of variation is the standard deviation expressed as a fraction of expected value.

6. Royston uses tail area error to find out which δ maximizes the parameter W of the Shapiro and Wilk's test of normality.

7. (1) may not necessarily lead to (14). Indeed, the simplest formulation generated by (1) is

$$x = V_1 \exp[(x_1 - \sigma_1 z_1 - \ldots)]$$

rather than (14). Since the compounding effect is now influenced by the standard deviation of z, not just by size, this variable is non proportional and cannot be used to form ratios. In case, however, of continuously compounding rates of change, then a proportional mechanism is obtained.