Fuzzy linear regression-based detection of earnings management

Henrik Höglund *

Hanken School of Economics, Handelsesplanaden 2, 65101 Vasa, Finland

ARTICLE INFO

ABSTRACT

A large number of accounting studies have examined the occurrence of earnings management in various contexts. In most of these studies, the earnings management detection model is based on the linear regression model suggested by Jones (1991). A considerable problem with the Jones model is the requirement of long time series of financial statement data. An alternative to estimating the linear regression model coefficients with ordinary least squares (OLS) is to use fuzzy linear regression (FLR) instead. One of the main advantages with FLR described in the literature is its ability to handle small data sets. The purpose of this study is to compare the performance of the OLS-based Jones model with the performance of the FLR-based Jones model. The results show that the performance of both types of models decreases when the length of the time series decreases and that there is no significant difference in the estimated discretionary accruals between the models. The results also show that the FLR-based Jones model outperforms the OLS-based Jones model in detecting simulated earnings management when the estimation time series is short. Overall, the results show that the FLR-based Jones model is a feasible alternative to the OLS-based Jones model, especially when the length of the estimation time series is restricted by data availability.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

A large number of accounting studies have examined the occurrence of earnings management in various contexts. There is, for example, evidence suggesting that firms manage earnings in order to avoid violating debt covenants (DeFond & Jiambalvo, 1994) or to influence initial public offering (IPO) valuations (Teoh, Welch, & Wong, 1998). A major challenge in earnings management studies is to find a measure for how much firms have managed their earnings. A common assumption is that earnings are managed through accounting accruals. With this assumption the measure of earnings management is the unexpected part of a firm's total accruals. Various models have been suggested for dividing the total accruals into non-discretionary (expected) and discretionary (unexpected) accruals. Most of these models are based on the earnings management detection model suggested by Jones (1991). The Jones model is a linear regression model where the level of total accruals is assumed to be explained by property, plant and equipment and the change in sales. In the original Jones model, the regression coefficients are estimated using a firm specific time series comprising data prior to the event year. The expected level of accruals is then calculated using the estimated coefficients with event year data.

A considerable problem with the Jones model is the requirement of long time series of financial statement data. Typically, in studies where the Jones model is used the requirement is set to at least ten years of data prior to the event year (e.g. Dechow, Sloan, & Sweeney, 1995; Thomas & Zhang, 2000). This requirement might lead to several problems, such as survivorship bias and non-stationary regression coefficients (Peasnell, Pope, & Young, 2000; Young, 1999). An alternative to estimating the firm specific regression coefficients with ordinary least squares (OLS) regression is using fuzzy linear regression (FLR), first introduced by Tanaka, Uejima, and Asai (1982). Contrary to probability theory-based OLS regression, FLR is based on possibility theory and fuzzy set theory. With FLR the objective is to minimize the fuzziness of the model rather than the sum of squared residuals. One of the main advantages with FLR described in the literature is its ability to handle small data sets (Shapiro, 2004). This was corroborated empirically by Kim, Moskowitz, and Koksalan (1996) who provided evidence that FLR outperforms statistical linear regression in predictive capability with small data sets. Thus, it is possible that the Jones model regression could be run with shorter time series using FLR. If this is the case, an FLR-based Jones model could reduce several problems originating from the requirement of long time series.

The purpose of this study is to compare the performance of the OLS-based Jones model with the performance of the FLR-based Jones model. The focus is on the time series version of the Jones model and the comparison is made using different lengths of the time series.

The remainder of this study is organized as follows. The basic operating principle of the linear regression-based accrual models is covered in Section 2. In Section 3 the estimation of FLR

* Tel.: +358 (0)40 3521768.
  E-mail address: henrik.hoglund@hanken.fi

coefficients and prediction with FLR models are covered. The research design is presented in Section 4 and the results from the empirical study are presented in Section 5. Section 6 concludes the study.

2. Time series-based discretionary accrual estimation models

During the past 30 years a large number of studies have examined the occurrence of earnings management in various contexts. Typically, the assumption in these studies has been that deviations from a normal or expected level of accruals constitute earnings management. In the early earnings management detection models the expected level of accruals was defined as the average accruals over an estimation period (Healy, 1985) or as the previous year accruals (DeAngelo, 1986). The shortcoming of these models is that they do not consider changing circumstances of the firms. For example, a normal increase in accruals as a result of an increase in sales would turn up as earnings management. Jones (1991) relaxed the assumptions that accruals remain stationary over time by suggesting a regression approach where the level of total accruals (*TACC*) is explained by the change in sales (Δ*REV*) and property, plant and equipment (*PPE*). The change in sales is assumed to explain the current accruals, such as receivables, inventory and payables, whereas property, plant and equipment should mainly explain the level of depreciation.

$$\frac{TACC_t}{TA_{t-1}} = \alpha_0 \frac{1}{TA_{t-1}} + \alpha_1 \frac{\Delta REV_t}{TA_{t-1}} + \alpha_2 \frac{PPE_t}{TA_{t-1}}$$

Typically, the variables in the Jones model are deflated with lagged total assets (TA) in order to reduce heteroscedasticity. The coefficients $\alpha_0 \ldots \alpha_2$ in the Jones model are estimated using a sufficiently long firm specific time series. The coefficients are then used to calculate the expected level or the non-discretionary part of total accruals using event period data. The difference between the observed event period total accruals and the calculated non-discretionary accruals is considered as unexpected accruals or discretionary accruals equaling earnings management. Jones (1991) used 13 years of data for estimating the regression coefficients whereas subsequent studies have typically settled for a minimum of 10 years (e.g. Dechow et al., 1995; Thomas & Zhang, 2000). A firm specific time series this long might lead to several problems. First, issues with survivorship bias are likely to arise (Bartov, Gul, & Tsui, 2000; Peasnell et al., 2000; Young, 1999). Second, a large number of firms do not have data stretching back ten years, reducing the size of the data set. Third, it is unlikely that the regression model coefficients remain stationary over a long period of time (Peasnell et al., 2000). Finally, the self-reversing nature of accruals can lead to problems with serial correlation of residuals (Peasnell et al., 2000).

To remedy some of these problems and shortcomings, DeFond and Jiambalvo (1994) suggested a cross-sectional approach where the regression coefficients are industry and year specific rather than firm specific. This approach does, however, also have its shortcomings. First, the industry membership is usually defined at a 2-digit SIC level. This might result in that some industries do not have a sufficient number of observations to run the regression, reducing the size of the data set. Second, the assumption is that the accrual generating process is similar among the firms in the same industry. There is, however, evidence suggesting that this assumption does not necessarily hold (Ecker, Francis, & Olsson, 2011; Höglund, 2013; Kothari, Leone, & Wasley, 2005). For example, Ecker et al. (2011) showed that the performance of the Jones model improved when running the regression with firms matched on lagged total assets rather than on industry membership.

3. Fuzzy linear regression

The concept of fuzzy linear regression (FLR) builds on the fuzzy set theory developed by Zadeh (1965) and it was first introduced by Tanaka, Uejima, and Asai (1982). The idea behind FLR is that the deviations between the observed and estimated values are assumed to originate from imprecise observations or vague relations between the model variables. This is contrary to OLS regression where random errors are assumed to be the reason for differences between observed and estimated values. Thus, the uncertainty in FLR models is fuzziness rather than randomness (Yang & Lee, 2002).

3.1. Estimating fuzzy linear regression coefficients

In OLS regression the objective is to minimize the sum of squared residuals. In FLR, on the other hand, the objective is to minimize the fuzziness of the model. The general form of the FLR model is:

$$\widehat{Y}_i = \tilde{A}_{0i} + \tilde{A}_{1i}x_{1i} + \cdots + \tilde{A}_{ji}x_{ji}$$

where $x_1 \ldots x_j$ denotes the independent variables, $\tilde{A}_0$ the estimated fuzzy intercept coefficient, $\tilde{A}_1 \ldots \tilde{A}_j$ the estimated fuzzy slope coefficients and $\hat{Y}_i$ the estimated fuzzy output. The fuzzy coefficients are generally represented as symmetric triangular fuzzy numbers where α_j equals the center value of $\tilde{A}_j$ and c_j the spread (see Fig. 1). However, other than symmetric triangular membership functions, such as asymmetric triangular and trapezoidal (Ishibuchi & Nii, 2001), are also used.

Before estimating the coefficients of the FLR, the fit between the FLR model and the data set has to be defined. This is done by setting the value of the h term, also called the target degree of belief (e.g. Chang & Ayyub, 2001), between 0 and 1. Each observed fuzzy $\tilde{Y}_i$ or crisp Y_i output must fall within the estimated fuzzy output $\hat{Y}_i$ at h (see Fig. 2). As the value of h increases, the fuzziness of the FLR model also increases. The h value does, however, not affect the center value of the fuzzy coefficients (Tanaka & Watada, 1988).

The FLR coefficients $\tilde{A}_0 \ldots \tilde{A}_j$ are estimated using linear programming with the objective of minimizing the total spread or fuzziness of the FLR model. The following objective function, where S denotes the total fuzziness of the regression model, m the number of independent variables and n the number of observations, is to be minimized:

$$S = \sum_{j=0}^{m} C_j \sum_{j=0}^{n} |x_{ji}|$$

The FLR linear programming problem comprises two sets of constraints. First, the spread c_j of the fuzzy coefficients $\tilde{A}_j$ has to be zero or positive. Second, all observed output variables $\tilde{Y}_i$ must fall within the estimated fuzzy output variables $\hat{Y}_i$ at level h. Y_i is the center
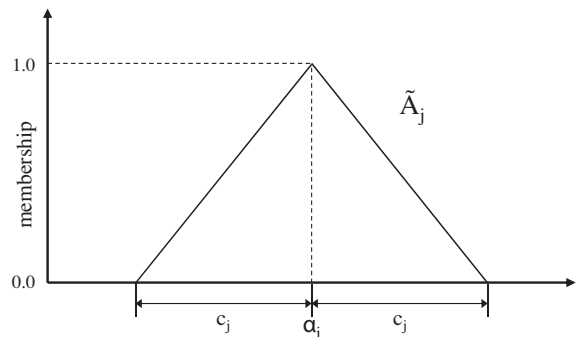

Fig. 1. Fuzzy coefficient with center value and spread.