Predicting Future Earnings Change Using Numeric and Textual Information in Financial Reports

Kuo-Tay Chen[1,*,**], Tsai-Jyh Chen[2], and Ju-Chun Yen[1]

[1] Department of Accounting, College of Management,
National Taiwan University
ktchen@management.ntu.edu.tw
[2] Department of Risk Management and Insurance,
National Chengchi University

Abstract. The main propose of this study is to build a more powerful earning prediction model by incorporating risk information disclosed in the textual portion of financial reports. We adopt the single-index model developed by Weiss, Naik and Tsai as a foundation. However, other than the traditionally used numeric financial information, our model adds textual information about risk sentiment contained in financial reports. We believe such a model can reduce specification errors resulting from pre-assuming linear relationship, thus can predict future earnings more accurately. The empirical results show that the modified model does significantly improve the accuracy of earning prediction.

Keywords: Single-index model, earnings prediction, risk sentiment, textual information.

1 Introduction

The neoclassical security valuation model determines a firm's value as the present value of expected future dividends. The ability of a firm to distribute dividends in the future can be assessed by its expected future earnings. As a result, future earnings prediction has become an important research issue. A number of studies have employed various factors in their models to predict earnings. These factors include bottom-line number of income statement (e.g. time-series pattern of earnings)[2], components of earnings [8], and accounting ratios based on income statement and balance sheet [13]. These studies utilize only numeric information in financial reports and do not incorporate textual information in their models. Since textual information such as footnotes and management discussion and analysis (MD&A) contain lots of information related to future earning, these models might have reduced their prediction power by excluding textual information.

Previous studies show that managers may have incentives to disclose risk in order to reduce the influence of future earnings shock and to avoid litigation responsibilities

* Corresponding author.
** Corresponding address: 1 Roosevelt Road, Sec. 4, Taipei, Taiwan, 106.

H. Chen et al. (Eds.): PAISI 2009, LNCS 5477, pp. 54–63, 2009.

[14]. Recently Li (2006) [12] uses risk-related words in financial reports as a measurement of risk sentiment and finds that risk sentiment is associated with future earnings and can be used to predict the change of earnings. This implies that risk sentiment in the textual part of financial reports may have information content about future earnings. As a result, our study builds an earning prediction model by incorporating the risk sentiment information contained in financial reports. Moreover, we do not assume that the risk sentiment has linear relation with future earnings. Because the disclosure of risk is manager's decision, future earnings may not definitely decrease when managers disclose low risk. On the other hand, large decrease of future earnings may be expected when high risk is disclosed. In other words, the relationship between risk sentiment and future earnings cannot be clearly specified as a linear relationship. Therefore, we might create a specification error if we pre-assume linear regression model without knowing the true relationship.

In this paper, we construct the research model based on Weiss, Naik and Tsai (2008) [16]. They employed a new method, single-index model, to estimate the market expectations of future earnings change (called Market-Adapted Earnings, MAE in [16]) and used MAE to predict future earnings. Compared with previous studies, single-index model allows a nonlinear relation between dependent and independent variables. Hence it can reduce the specification error committed by the pre-assumed linear model. Weiss et al. [16], however, consider merely numeric accounting ratios while assessing MAE. As we mentioned before, the textual part of financial reports may also contain useful information for market participants to predict future earnings. For this reason, we adjust Weiss et al. [16] model by incorporating the textual portion of financial reports to re-examine MAE, and hope to build an earnings predicting model with more predicting power.

In summary, we design our study in two parts. First, we construct a single-index model that includes a risk sentiment variable which represents textual information in financial reports. From this model we assess MAE to predict future earnings change. Second, we calculate predicted market expectation of future earnings change (predicted MAE) and compare it with those generated by Weiss et al. predicting model to determine the relative strength of our model.

2 Literature Review and Hypothesis Development

In this section we first review previous studies on earnings prediction and textual information. We then develop our research hypotheses.

2.1 Earnings Prediction

2.1.1 Factors to Predict Earnings

In 1970s, several studies try to find out the time-series pattern of earnings (e.g. Ball and Watts 1972 [2]). Those studies suggest that earnings process is close to random walk with a drift. Beaver, Lambert and Morse (1980) [3] assert that earnings are a compound process of the earnings series which will reflect events affecting price and those which will not affecting price. Those studies use the bottom-line earnings number to predict

future earnings. Brown et al. (1987) [5] compare the quarterly earnings predicting results of three different time-series earnings model and analyst earnings forecast from Value Line. Among one-quarter ahead through three-quarter ahead earnings predicting results, however, all of the three time-series earnings model do not outperform than analyst forecasts by examining forecast error.

After that, some researchers try to construct more specific and accurate earnings predicting models. Rather than using bottom-line earnings, Fairfeild et al. (1996) [8] suggest that different components of earnings may contain different content to predict earnings. They disaggregate earnings into several components in different way to see which approach of classification will provide the best ability of predicting one-year ahead return on equity. The result shows that non-recurring items of earnings do not have better predicting ability than recurring items. In addition, their model performs better than the model with aggregated bottom-line item.

Besides, since the time-series of earnings model only extract the information from income statement, several studies turn to exam the content of the balance sheet items and accounting ratios to predict future earnings. Ou (1990) [13] uses non-earnings annual report numbers to predict earnings change. In this study, logit model is used to predict the sign of earnings change in next period. Lev and Thiagarajan (1993) [11] choose 12 accounting ratios as candidates and calculate an aggregate fundamental quality score. First they assign 1 to positive signal and 2 to negative signal for each of the 12 accounting ratios, and then compute an average score for each firm and year. They find that this fundamental quality score has positive relation with subsequent earnings changes. Abarbanell and Bushee (1997) [1] employ an earnings predicting linear model with accounting ratios tested by Lev and Thiagarajan [11].

Other than accounting information, Beaver et al. (1980) [3] suggest that price reflect the information not presented in current earnings. As a result, price may also contain the information content and lead earnings. Beaver et al. (1987) [4] use percentage of price change as an independent variable to predict percentage of earnings changes.

2.1.2 Statistics Methods to Predict Earnings

The previous subsection focuses on the factors or elements which can be used to approximate future earnings. The other studies line is to adopt different research methods to improve the ability of predicting earnings. Most of the previous studies use linear OLS regression model (e.g. [3], [4], [8]), logit model (e.g. [13]). However, it is unknown that what the true pattern of relation between the predicting factors and future earnings is. For example, some have found S-shape relationship in returns-earnings relationship studies [6]. Therefore, the use of parametric estimation model may result in the problem of specification error.

The main propose of Weiss et al. (2008) [16] is to develop a new index to extract forward-looking information from security price. This study originates from the price lead earnings researches of Beaver et al. who suggest that price may have information content about future earnings. Thus, Weiss et al. take a semi-parametric statistic method, single-index model, to connect the relation between earnings change and returns via market expected earnings change. It extracts information from both security prices and accounting ratios to calculate a new index representing market's expectations of future earnings change, which they called MAE. The use of single-index model

allows for a nonlinear returns-earnings relationship. They use earnings change and four accounting ratios such as change in inventory, change in gross margin, change in sales and administrative expenses and change in accounts receivable as independent variables and annual returns as dependent variable. This paper has two contributions: (1) the single-index model performs more explanatory power than the linear model with fundamental accounting signals by Lev and Thiagarajan (1993)[11], and (2) the predicted MAE index, which is used to predict future earnings change, has better forecasting ability than previous random-walk model and accounting-based forecasting model by Abarbanell and Bushee (1997) [1]. But still, MAE index does not outperform analyst earnings forecast from I/B/E/S.

2.2 Textual Information

Li (2006) [12] examines the relation between risk sentiment in annual reports and future earnings. Li measures risk sentiment by counting frequency of words about risk or uncertainty in textual part of annual reports. The counting rules in his paper are (1) count the frequencies of the "risk" words (including "risk", "risks" and "risky") and the "uncertainty" words (including "uncertain", "uncertainty" and "uncertainties"). (2) "risk-" format words are excluded because it may relate to "risk-free". Li finds a negative relation between risk sentiment and next period earnings.

2.3 Hypothesis Development

After MAE has been estimated, we can use predicted MAE to forecast future earnings. Because the risk sentiment in financial reports may affect future earnings rather than current earnings, it may have information content to future earnings. That means, the change of risk sentiment in financial reports may add incremental earnings prediction ability. Accordingly, the hypothesis is developed as follows:

Hypothesis: The earnings forecast errors of the modified model with risk sentiment variable are lower than previous models.

3 Research Design

3.1 Introduction to Single-Index Model

As we mentioned in previous sections, using parametric method when unknowing the true pattern may cause specification errors. Although nonparametric method can be used to solve this problem, but the cost of reducing specification errors can be very high because (1) estimation precision will decrease when the dimension of independent variables increase, (2) it's hard to interpret the results in multidimensional independent variables, and (3) it cannot be used to predict [9]. Since that, semi-parametric method can both solve the problem from nonparametric methods and reduce the specification errors from parametric methods.

Single-index model is one of the semi-parametric models. It aggregates the multi-dimensional X into single-dimensional index first, and then estimates the function connecting the dependent variable and the index by parametric estimation methods. The basic form of single-index model is as follows:

$$Y_i = G(X\beta) + \varepsilon_i \tag{1}$$

where â is the vector of coefficients and *X*â is the index. Note that we do not have to assume the type of G(.) in priori. G(.) can be an identity function (then the function (1) becomes linear model), cumulative distribution function (then the function (1) becomes probit model) or nonlinear function. In turn, G(.) is determined endogenously.

In this paper, we employ the same estimating method with Weiss et al. [16] as follows:

1. Estimating â: There are different approaches can be adopted to estimate â(e.g. nonlinear least square), even without knowing G(.). One simple method without solving optimal problem is sliced inverse regression [7]. First the data should be sorted by the increasing value of dependent variable, and then divided into several groups, or slices. After slicing, calculate a new covariance matrix by slice means and then estimateâ. This method is without solving optimal problem and link-free. That is, â can be estimated when G(.) is unknown.
2. Estimating G(.): Afterâis estimated, the index *X*â can be calculated. As a result, we can use *Y* and *X*â to estimate G(.) by nonparametric method since the multidimensional *X* has been aggregated into single dimensional index.

3.2 Model Construction

3.2.1 Single-Index Model

Following Weiss et al. [16], we also release the relation pattern between returns and earnings by setting G(.) allowing linear or nonlinear relation. However, we want to examine whether the risk sentiment in the textual part of financial reports has information content to future earnings. Therefore, we incorporate additional variable to capture risk sentiment in 10-K reports in Weiss et al. model and construct as follows:

$$\begin{aligned} &R_{it} = G(MAE_{it}) + \varepsilon_{it} \\ &\text{Where } MAE_{it} = \Delta E_{it} + \beta_{1t}\Delta INV_{it} + \beta_{2t}\Delta GM_{it} + \beta_{3t}\Delta SGA_{it} + \beta_{4t}\Delta REC_{it} + \beta_{5t}\Delta RS_{it} \end{aligned} \tag{2}$$

– R_{it}: Annual abnormal returns, which is measured by accumulated 12 months of monthly raw stock returns starting from the fourth month of the beginning of fiscal year to the fourth month of the ending of fiscal year, and then less the equally weighted monthly returns for the same periods in CRSP [16].

– ΔE_{it}: Change of earnings per share before extraordinary items and discontinued operations, deflated by the stock price at the beginning of fiscal year. [16] [10] Note that we set the coefficient of ΔE_{it} equals one for scale normalization. [9]

- ΔINV_{it} : Change in inventory measured by $\Delta Inventory - \Delta Sales$ [1], which is a signal of logistic operations[2]. (Compustat #78 or 3, #12)
- ΔGM_{it} : Change in gross margin measured by $\Delta Sales - \Delta GrossMargin$, which is a signal of profitability of sales. (#12,#12-41)
- ΔSGA_{it} : Change in sales and administrative expense measured by $\Delta SalesAndAdministrativeExpense - \Delta Sales$, which is a signal of marketing and administrations. (#189, #12)
- ΔREC_{it} : Change in accounts receivable measured by $\Delta AccountsReceivable - \Delta Sales$, which is a signal of management of clientele.
- ΔRS_{it} : Change in risk sentiment in the MD&A and footnote parts of annual report. This variable is used to capture the annual reports' information which will affect future earnings but not recognized in financial statements yet. In other words, this variable presents textual information rather than numerate information in annual reports. Following Li [12], ΔRS_{it} is calculated as follows:

$$\Delta RS_{it} = \ln(1 + NR_{it}) - \ln(1 + NR_{it-1}) \tag{3}$$

Where NR_{it} are the numbers of occurrences of risk related words in the annual report of year *t*. The risk related words are the words including "risk" (e.g. Risk, risks, risky) and "uncertainty" (e.g. Uncertain, uncertainty, uncertainties) and excluding "risk-".

Note that based on the definitions of the above variables and the analysts' interpretation, it was defined as "good news" when the value of the variable is negative, and vise versa. As a result, we predict the signs of all the coefficients are negative, including ΔRS_{it} .

3.2.2 MAE Prediction

Since we want to compare the predicting ability of our model with Weiss et al. model, we adopt the same procedure with Weiss et al. After theβare estimated, estimated *MAE* can be calculated: [16]

$$\begin{aligned} \hat{MAE}_{it} &= \Delta E_{it} + \sum_j S_{jit}\hat{\beta}_{jt} \\ &= \Delta E_{it} + \hat{\beta}_{1t}\Delta INV_{it} + \hat{\beta}_{2t}\Delta GM_{it} + \hat{\beta}_{3t}\Delta SGA_{it} + \hat{\beta}_{4t}\Delta REC_{it} + \hat{\beta}_{5t}\Delta RS_{it} \end{aligned} \tag{4}$$

[1] Ä means "percentage change of the variable between its actual amount and expect amount, where the expected amount is the average amount in the previous two year. E.g. $E(Sales_t) = (Sales_{t-1} + Sales_{t-2})/2$, $\Delta Sales = [Sales_t - E(Sales_t)]/E(Sales_t)$. See Lev and Thiagarajan (1993) [11].

[2] Originally we should use change in cost of goods sold as a benchmark rather than change in sales. But analysts usually use change in sales and previous study showed the identical results [11]. In order to compare with previous studies, we also choose change in sales as benchmark.

Next, in order to estimate future earnings, the estimated MAE is transformed as follows[3]: [16]

$$\hat{MAE}_{it}^{*} = \hat{MAE}_{it} \times \frac{\overline{\Delta E_{it}}}{\overline{\hat{MAE}_{it}}} = \frac{\hat{MAE}_{it}}{\hat{MAE}_{it}} \times \overline{\Delta E_{it}} \tag{5}$$

Where $\overline{\Delta E_{it}}$ and $\overline{\hat{MAE}_{it}}$ are the means of actual earnings change and estimated MAE. After transforming, the mean of $\hat{MAE}_{it}$ will equal the mean of ΔE_{it}.

4 Empirical Results

4.1 Data

To measure the fundamental accounting ratios, we use data from Compustat fundamental annual and calculate the percentage change of inventory, gross margin, sales and administrative expense, and receivables. In addition to the accounting ratios, we use earnings per share before extraordinary items in Compustat to calculate the change of reported earnings, which is deflated by the share price at the beginning of the fiscal year.

For abnormal stock returns, we use data of raw stock monthly returns from Compustat and equally weighted monthly return from CRSP. Abnormal annual stock returns are calculated by cumulating raw stock monthly returns minus equally weighted monthly return for 12 months (from the fourth month after the beginning of fiscal year to the third month after the end of the fiscal year) [16].

Following Li, we extract number of risk related words in 10-K reports to calculate change of risk sentiment. The counting method of risk related words is similar to Li[4].

The data period is from 1998 to 2006. The sample firm with missing data will be dropped for that sample year.

4.2 Earnings Forecast Error

In Weiss et al. 2008 paper, the predicting results of this model has been proved outperforming than random-walk model and fundamental accounting ratios model proposed by Abarbanell and Bushee (1997) [1]. Accordingly, we can only compare our model, which contains textual information in financial reports, with Weiss et al. model by median absolute forecast errors. Our purpose is to see whether risk sentiment in 10-K reports can improve earnings predicting ability; therefore, we compare earnings

[3] In Weiss et al., they state that transforming can let the sum of noise in earnings change be zero across firms. Moreover, transformed MAE is equivalent to random-walk model. [16].

[4] Different from Li, we do not delete the title items in 10-K reports before counting risk words. However, we expect similar results since we use changes of risk words but no absolute number of risk words for the year.

predicting errors of SIM model with change of risk sentiment variable with those of the original SIM model in Weiss et al.

To compare the predicting abilities among the models, absolute earnings forecast errors are calculated: [16]

Absolute earnings forecast error =

$$\left|\frac{Actual - Forecastd}{Actual}\right| = \left|\frac{\Delta E_{t+1} - Forecasted\,\Delta E_{t+1}}{E_{t+1}}\right| = \left|\frac{\Delta E_{t+1} - M\hat{A}E^{*}_{t+1}}{E_{t+1}}\right| \quad (6)$$

After absolute earnings forecast errors are calculated for the two models, median of the forecast errors is reported rather than mean forecast errors to prevent the influence of huge value owing to deflation.

In traditional linear earnings forecast model [1], earnings in period t and fundamental ratios in period t-1 are used to construct model, and then we can input fundamental ratios in the model to predict earnings in period t+1. That is, we need two year data to predict next year earnings. However, SIM model needs only one year data, rather than two years, to construct MAE*. Therefore, we estimate next year's earnings change using both the model construct by last year's data (out of sample prediction) and the model construct by this year's data (in sample prediction). The results are showed in table 1 and table 2.

When using last year's model and this year's data (out-of-sample prediction), incorporating additional risk sentiment reduce median forecast error by nearly 60% in full sample (from 0.146244 to 0.058998). In addition, the nine year average of median forecast error also declines by 70% (from 0.36856 to 0.10973).

Table 1. Median forecast errors of MAE* with and without use of risk sentiment. (out-of-sample)

Year	Median forecast error	
	MAE* with no risk sentiment	MAE* with risk sentiment
1998	--	--
1999	0.068852	0.082984
2000	0.065603	0.065896
2001	0.159866	0.056722
2002	1.477695	0.243611
2003	0.423893	0.364171
2004	0.722144	0.033904
2005	0.011548	0.011429
2006	0.018875	0.01912
Average	0.36856	0.10973
Full sample	0.146244	0.058998

Table 2. Median forecast errors of MAE* with and without use of risk sentiment. (in sample)

Year	Median forecast error	
	MAE* with no risk sentiment	MAE* with risk sentiment
1998	0.040108	0.015707
1999	0.073343	0.074516
2000	0.064788	0.071767
2001	0.027909	0.027053
2002	0.122653	0.119972
2003	0.687346	0.18015
2004	0.047517	0.035202
2005	0.011333	0.011427
2006	0.034801	0.077692
Average	0.123311	0.068165
Full sample	0.061508	0.055131

When using this year's model and data to prediction next's year's earnings change (in sample prediction), considering additional risk sentiment reduce median forecast error by 10% (from 0.061508 to 0.055131) in full sample. The nine year average of medina forecast error is reduced by 45% (from 0.123311 to 0.068165). Comparing to the results of out-of-sample prediction, the percentages of reduction are less than out-of-sample prediction, but the absolute values of forecast error in in-sample prediction are smaller.

As a result, the use of risk sentiment in textual part of annual reports, additional to numeric information in financial statement, can improve the earnings predicting ability.

5 Conclusion

Many studies have attempted to make better predictions of earnings. Nevertheless, numeric information may only provide partial information for future earnings. Thus, in this study we incorporate textual information in financial reports to examine whether it has incremental earnings prediction ability. The results show that incorporating risk sentiment in 10-K reports can significantly improve the one-year ahead earnings prediction ability by using a single-index model.

References

1. Abarbanell, J.S., Bushee, B.J.: Fundamental Analysis, Future Earnigns, and Stock Returns. Journal of Accounting Research 35, 1–24 (1997)
2. Ball, R., Watts, R.: Some Time Series Properties of Accounting Income. Journal of Finance, 663–682 (June 1972)

3. Beaver, W., Lambert, R., Morse, D.: The Information Content of Security Prices. Journal of Accounting and Economics 2, 3–28 (1980)
4. Beaver, W., Lambert, R.A., Ryan, S.G.: The Information Content of Security Prices: a Second Look. Journal of Accounting and Economics 9 (1987)
5. Brown, L.D., Hagerman, R.L., Griffin, P.A., Zmijewski, M.E.: Security Analyst Superiority Relative to Univariate Time Series Models in Forecasting Quarterly Earnings. Journal of Accounting and Economics 9, 61–87 (1987)
6. Das, S., Lev, B.: Nonlinearities in the Returns-Earnings Relation: Tests of Alternative Specifications and Explanations. Contemporary Accounting Research 11, 353–379 (1994)
7. Duan, N., Li, K.C.: Slicing Regression: A Link-Free Regression Method. Annals of Statistics 19, 503–505 (1991)
8. Fairfield, P.M., Sweeney, R.J., Yohn, T.L.: Accounting Classification and the Predictive Content of Earnings. The Accounting Review 71(3), 337–355 (1996)
9. Horowitz, J.L.: Semiparametric Methods in Econometrics. Springer, New York (1998)
10. Kothari, S.P.: Price-Earnings Regressions in the Presence of Prices Leading Earnings. Journal of Accounting and Economics 15, 143–171 (1992)
11. Lev, B., Thiagarajan, R.: Fundamental Information Analysis. Journal of Accounting Research 31, 190–215 (1993)
12. Li, F.: Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports? Working paper (2006)
13. Ou, J.A.: The information Content of Nonearnings Accounting Numbers as Earnings Predictors. Journal of Accounting Research 28, 144–163 (1990)
14. Skinner, D.J.: Why Firms Voluntarily Disclose Bad News. Journal of Accounting Research 32, 38–60 (1994)
15. Wang, T.W., Rees, J.: Reading the Disclosures with New Eyes: Bridging the Gap between Information Security Disclosures and Incidents. Working paper (2007)
16. Weiss, D., Naik, P.A., Tsai, C.L.: Extracting Forward-Looking Information from Security Prices: A New Approach. The Accounting Review 83, 1101–1124 (2008)