

**A BAYESIAN APPROACH TO VARIABLE SELECTION
IN LOGISTIC REGRESSION WITH APPLICATION TO
PREDICTING EARNINGS DIRECTION FROM
ACCOUNTING INFORMATION**

RICHARD GERLACH, RON BIRD AND ANTHONY HALL

School of Finance and Economics
University of Technology, Sydney
Sydney NSW 2007
Australia

`richard.gerlach@uts.edu.au`

`ron.bird@uts.edu.au`

`tony.hall@uts.edu.au`

ABSTRACT. This paper presents a Bayesian technique for the estimation of a logistic regression model including variable selection. The model is used, as in Ou and Penman (1989), to predict the direction of company earnings, one year ahead of time, from a large set of accounting variables from financial statements. We present a Markov chain Monte Carlo sampling scheme, that includes the variable selection technique of Smith and Kohn (1996) and the non-Gaussian estimation method of Mira and Tierney (1997), to estimate the model. The technique is applied to companies in the United States, United Kingdom and Australia. This extends the analysis of Ou and Penman (1989) who studied United States companies only. The results obtained compare favourably to the technique used in Ou and Penman (1989) for all three regions.

KEYWORDS : Slice sampler, Stepwise regression

1. INTRODUCTION

Ou and Penman (1989) use a logistic regression model to predict the direction of company earnings one year ahead of time using financial statement items. They employ an investment strategy for United States (US) companies from 1973-1983 based on forecasts from this model. The regressors included in the model are selected from a set of 68 variables via a process known as stepwise regression. This procedure is simple to understand, easy to compute and is thus widely used. However, statistically speaking this procedure suffers from the following drawbacks. Firstly, the variables selected are not chosen to maximise the likelihood of the model, nor any other reasonable optimization criterion. In addition, the estimated significance of each variable in this process is biased and may be largely inflated or deflated. These drawbacks are well known. See Weisburg (1985) for a discussion of stepwise regression.

The use of Markov chain Monte Carlo (MCMC) methods has allowed the Bayesian estimation of complex models unable to be estimated via classical techniques such as maximum likelihood. Bayesian methods set parameter estimates to the modes of the pseudo-likelihood of the model. This pseudo-likelihood is the product of the likelihood and the prior distribution placed on the model. The choice of diffuse priors allows maximum likelihood estimators to be produced. Smith and Kohn (1996) illustrate how to use MCMC methods to perform variable selection in a simple linear regression setting. We incorporate this technique into a MCMC sampling scheme that performs variable selection for a logistic regression model. The technique is illustrated by performing a similar analysis to that by Ou and Penman (1989), and then comparing with the results obtained by stepwise regression.

The logistic regression model involves a binary response variable and hence is not a Gaussian model. Mira and Tierney (1997) propose the slice sampler as a general technique for the Bayesian estimation of statistical models. Gerlach and Kohn (1998) illustrates the use of the slice sampler for non-Gaussian state space models with Gaussian transition equations. A logistic regression is one case of such a model and we use the version of the slice sampler in Gerlach and Kohn (1998) as part of the estimation procedure.

The paper is presented as follows. Section 2 discusses the data and the model. Section 3 presents the Bayesian estimation procedure used and reviews the stepwise regression procedure used in Ou and Penman (1989). Section 4 presents and discusses the results and compares the two procedures for the US, United Kingdom (UK) and Australia.

2. DATA AND MODEL

2.1. Discussion of data. Ou and Penman (1989) obtain annual financial statement information from US companies in 1973-1983 from the 1984 COMPUSTAT annual report files. They compute a measure of company earnings surprise by comparing change in earnings per share (EPS) with a forecasted change in EPS figure. This forecast is the average of the previous four years of change in EPS. If this earnings surprise measure is positive they record an observation of 1, otherwise a 0 is recorded. These form the observations of a logistic regression model. The potential regressors in the model are 68 accounting variables in the COMPUSTAT database. Using the COMPUSTAT 1998 North American files we were able to obtain information on 63 of these variables in the years 1982-1997, including their research database to capture non-continuing companies. This eliminates any bias in

the study from including only companies that have survived the whole sample period. Further, the companies included are all non-financial companies with market capitalizations in excess of *US*100 million dollars as at the end of 1998. This minimum capitalization figure was applied to ensure that there is sufficient liquidity to support any strategy that might be developed and it was adjusted each year in line with movements in Russells 3000. Typically, after deleting companies with missing observations this left around 1700 companies to be analysed each year.

Using the Global Vantage 1998 International files we obtained information on 52 variables in the UK from 1990-1997. After missing observations were deleted around 400 companies were analysed each year. In addition information on 47 variables was obtained for Australia from 1990-1997, with around 225 companies included each year.

As in Ou and Penman (1989) we used a five year window to select the significant variables and estimate the corresponding regression coefficients for both techniques, i.e. the MCMC technique presented in section 3 and the stepwise regression procedure. One measure we use to compare the two variable selection techniques is the accuracy of predictions of earnings increase over this in-sample period. The models identified by each technique over the five year period are then applied to forecast only the next year of earnings increases. Another measure of comparison is then forecast accuracy over each year. The five year window is moved forward one year and the analysis is repeated including the estimation of new models. For the US this gives forecasts for each year from 1988-1997. For the UK and Australia forecasts were obtained for each of 1993-1998. Due to availability of UK and Australian data only a two year window was used to obtain the model used in

forecasting 1993, a three year window in 1994 and a four year window in 1995.

Many of the accounting variables used in this study are ratios or products of some kind. They are thus prone to outlying observations. For this reason, prior to the variable selection procedures being performed, many of the variables were transformed using either a square root or logarithmic transform.

2.2. Model. The model used by Ou and Penman (1989) is a dynamic logistic regression model, where the observations, y , record whether change in earnings per share is larger than its four year average ($y_t = 1$), or smaller ($y_t = 0$). The probability of an earnings increase is allowed to change over time. i.e.

$$(2.1) \quad P(y_t = 1) = \pi_t, \quad t = 1, \dots, n$$

The logarithm of the odds, i.e. the log-odds, of an earnings increase is modelled as a simple linear regression as follows.

$$(2.2) \quad \log\left(\frac{\pi_t}{1 - \pi_t}\right) = x_t = X_t\beta + e_t, \quad t = 1, \dots, n$$

where the errors e_t are normally distributed i.e. $e_t \sim N(0, \sigma^2)$. The row vector X_t represents the values for the accounting variables for observation t . These values are lagged by one year. For instance, when considering the earnings increases in the year 1996, the accounting variables used are those for the year 1995 and hence generally available in early 1996. This allows us to have the required accounting information actually available at the time when we make our one year ahead forecasts and hence avoids any look-ahead bias in these results.

3. ESTIMATION TECHNIQUES

3.1. Discussion. This section presents the MCMC estimation technique applied to the logistic regression model in (2.1-2.2). This technique is employed to select the accounting variables to be included in the model. The stepwise regression procedure employed in Ou and Penman (1989) is also reviewed. For both selection procedures, once the variables have been chosen a standard maximum likelihood analysis is run to estimate the model then predict and forecast the observations in the years analysed. Although the coefficient estimates and predictions can be obtained from the MCMC output, we want to test the variable selection technique against stepwise regression and not MCMC against maximum likelihood. We thus run a likelihood analysis for each method to avoid confounding what we are testing.

3.2. MCMC variable selection. Following Smith and Kohn (1996) auxiliary variables are introduced to indicate which accounting variables are to be included in the model. These variables are denoted by J , where $J_i = 1$ indicates that the i th variable is to be included in the model and $J_i = 0$ indicates the opposite. The goal of the estimation scheme is to estimate the posterior probability that each variable is to be included in the model. i.e. $P(J_i = 1|y)$ for $i = 1, \dots, M$. As mentioned earlier, for the US $M = 63$. This goal is achieved by integrating out the model parameters from the full conditional posterior distribution $P(J_i = 1|y, x, \beta, \sigma^2, J_{\neq i})$. The variables β and σ^2 can be integrated out analytically as shown below. The variables x and $J_{\neq i}$ are then numerically integrated out by iterating over the MCMC sampling scheme detailed below. An estimate of $P(J_i = 1|y)$ is obtained

for each variable via the following average

$$(3.1) \quad \hat{P}(J_i = 1|y) = \frac{1}{D} \sum_{j=1}^D P(J_i = 1|y, x^{[j]}, J_{\neq i}^{[j]})$$

where $x^{[j]}$ and $J_{\neq i}^{[j]}$ are the j th iterate estimates for these variables in the MCMC sampling scheme.

3.2.1. *Details of MCMC scheme.* The following section on calculating the above average is a summary of the work in Smith and Kohn (1996).

$$(3.2) \quad p(J_i|y, x, J_{\neq i}) \propto p(y|x)p(x|J)p(J_i|J_{\neq i}).$$

We set the prior $p(J_i|J_{\neq i}) = 0.5$ as a constant so that all variables are a priori equally likely to be included in the model. This prior could easily be changed to reflect any beliefs about certain variables being included, or not included, in the model. Then,

$$(3.3) \quad p(x|J) = \int \int p(x|\beta, \sigma^2, J)p(\beta, \sigma^2)d\beta d\sigma^2.$$

Following Smith and Kohn (1996) we allow the following priors. $\beta|\sigma^2 \sim N(0, c\sigma^2(X'X)^{-1})$ and $p(\sigma^2) \propto \frac{1}{\sigma^2}$. These are standard prior distributions allowing the variable $\log(\sigma^2)$ to have a diffuse or flat prior on the real line. We choose $c = 100$ to increase the variance of the prior for β and make it reasonably diffuse. As shown by Smith and Kohn (1996) the integral can then be performed analytically so that

$$(3.4) \quad p(x|J) \propto S(J_i)^{-\frac{q}{2}}(c+1)^{-\frac{q}{2}}$$

where $S(J_i) = x'x - B'A^{-1}B$, $B = X'x$, $A = \frac{(c+1)}{c}X'X$ and q is the number of accounting variables currently in the regression. It is then easy to show that

$$(3.5) \quad P(J_i = 1|y, x, J_{\neq i}) = \frac{1}{1 + h_i}$$

where

$$(3.6) \quad h_i = \left[\frac{S(J_i = 1)}{S(J_i = 0)} \right]^{n/2} \sqrt{1+c} \frac{p(J_i = 1|J_{\neq i})}{p(J_i = 0|J_{\neq i})}.$$

As mentioned previously we must also generate the vector of log odds of earnings increase, x , as part of the sampling scheme. We do this by generating from its full conditional posterior distribution as follows

$$\begin{aligned} p(x_t|y, x_{\neq t}, \beta, \sigma^2) &\propto \frac{e^{x_t y_t}}{1 + e^{x_t}} \exp \left[-\frac{1}{2\sigma^2} (x_t - X_t \beta)^2 \right] \\ &\propto \frac{1}{1 + e^{x_t}} \exp \left[-\frac{1}{2\sigma^2} (x_t - X_t \beta - \sigma^2 y_t^2)^2 \right]. \end{aligned}$$

Note that this distribution has a Gaussian part

$g(x_t) = \exp \left[-\frac{1}{2\sigma^2} (x_t - X_t \beta - \sigma^2 y_t^2)^2 \right]$ and a non-Gaussian part $l(x_t) = \frac{1}{1+e^{x_t}}$. These two parts when combined form an unknown distribution.

However, the two parts can be written as a non-Gaussian observation state space model with a Gaussian state transition equation. We can thus use the slice sampler as detailed in Gerlach and Kohn (1998). The auxiliary variable u is introduced so that $u_t|x_t \sim \text{Unif} \left[0, \frac{1}{l(x_t)} \right]$. Then we can generate each x_t as follows

$$(3.7) \quad x_t|y, x_{\neq t}, \beta, \sigma^2, u \sim N_{\text{cons}}(X_t \beta + \sigma^2 y_t^2, \sigma^2)$$

which is a normal distribution constrained so that $u_t < l(x_t)$ or $x_t < \log \left(\frac{1-u_t}{u_t} \right)$. The slice sampler thus transforms the non-Gaussian density into a constrained Gaussian density. Note that the uniform variable u is also generated as part of the MCMC sampling scheme. For a discussion of the slice sampler see Mira and Tierney (1997).

As the posterior distribution for each x_t depends on both β and σ^2 , we must also generate these variables as part of the sampling scheme. The posterior distributions for $\beta|y, x, J$ and $\sigma^2|y, x, J, \beta$ are easy to

compute, using the prior distributions detailed above, and turn out as a multi-variate t and an inverse gamma distribution respectively. i.e.

$$\beta|y, x, J \sim t_{n+q} \left(A^{-1}B, \frac{A^{-1}D}{n+q} \right)$$

$$\sigma^2|y, x, J, \beta \sim \text{Inv. Gamma} \left(\frac{n+q}{2}, \frac{C}{2} \right)$$

where $A = \frac{c+1}{c}X'X$, $B = X'x$, $C = (x - X\beta)'(x - X\beta) + \frac{1}{c}\beta'X'X\beta$ and $D = x'x - B'A^{-1}B$.

3.2.2. *Sampling Scheme.* Initial values are randomly chosen for the model parameters, β, σ^2, J and x . Iterates are successively generated in turn from each of the posterior distributions detailed above. The iterates for $P(J_i = 1|y, x^{[j]}, J_{\neq i}^{[j]})$ are saved and used in (3.1).

3.3. **Stepwise regression.** We follow the procedure in Ou and Penman (1989) as follows. For each of the variables considered (e.g. each of the 63 for the US database) a univariate maximum likelihood logistic regression analysis is run. Variables whose p -value for the estimated regression coefficient is less than 10 percent are included in the second stage while the rest are discarded. The second stage of the procedure involves a backward elimination stepwise regression procedure starting with all the variables that survived the first stage. This consists of successively fitting the logistic regression with all remaining variables and eliminating them one at a time using a 10 percent level. This continues until all variables remaining in the model are significant at the 10 percent level.

4. RESULTS

4.1. **US companies.** As described in section 2 the two variable selection techniques are used to estimate the model for US companies over

a moving five year window. The first window is 1983-1987, the next is 1984-1988 and the last is 1992-1996. This allows one year forecasts of earnings increase for the years 1988 up to 1997 inclusive. Note that the accounting variables in year $k - 1$ are used in the regression to predict earnings direction in year k . For example, this allows us to forecast earnings direction in 1997 using information available in 1996.

The results are summarised in tables 1-5. Table 1 shows the variables selected by each method for each year forecasted. Typically the Bayesian method chooses about 15 – 18 variables as significant in each year. The stepwise regression procedure has much more variation in the number of variables it chooses. For instance, in 1991 it chooses only 4 variables, whereas in 1994 it chooses 34 variables. Typically this method selects less variables than the Bayesian approach. The five year windows are overlapping, for instance the year 1992 is in five of these windows. Also, successive five year windows contain four common years. We would thus expect that many of the same variables would be selected from year to year and from successive windows. This persistence pattern in selection shows up clearly in the Bayesian variables chosen. The variables selected by stepwise regression appear much more sporadic with less pattern over the years. This inconsistency in variables selected by stepwise regression may illustrate that the likelihood is not being maximised by this procedure and that the 'best' variables are not being chosen.

Tables 2 and 3 contain the results showing the in-sample accuracy of both methods over each of the five year windows considered. For example, for the years 1992-96 when the model selected by MCMC estimated the probability of earnings increase as being greater than 0.8, earnings did actually increase (i.e. $y=1$) 88 percent of the time. Most years seem

to be comparable for the two techniques. Note that we would expect the Bayesian method to perform as well if not better than the stepwise regression method and this seems to be the case. The column labelled \bar{y} contains the actual percentage of earnings increases in each five year window. Comparing the results with this column shows clearly that the models estimated are capturing useful information about earnings direction. In other words the models do much better than just guessing.

The last row in tables 2 and 3 contains a weighted average of each column with the number of observations used to form each percentage, used as the weights. This is done because, for example, the number of observations with Pr above 0.8, say, varies a lot from year to year, so a weighted average is a better summary measure than a mean. The averages are almost exactly the same for each method, with the Bayesian method doing marginally better, especially when Pr is very high or very low. On average, when Pr is above 0.6 or below 0.4 the Bayesian method is accurate in prediction 70 percent of the time and the stepwise regression procedure is 69 percent accurate. These are the values that Ou and Penman (1989) base their investment strategies upon, that is they invested in those companies whose value of Pr was above 0.6 and they bet against those companies whose values of Pr were below 0.4.

Tables 4 and 5 show the forecast accuracies for each year from 1988-1997. Again the numbers are very comparable for each method. However, on average the Bayesian technique again does marginally better, especially when Pr is very high or very low. When Pr is above 0.6 or below 0.4 the Bayesian method is accurate in forecasting 67 percent of the time and the stepwise regression procedure is 64 percent accurate.

It is interesting to note how similar the two methods perform in accuracy and compare this with the different models selected by each

in table 1. Even if the stepwise regression procedure does not maximise the likelihood it still performs quite well in comparison to the statistically sound Bayesian technique.

4.2. UK companies. The results for the UK are summarised in tables 6-10. Table 6 presents the variables selected by each method. Once again we note that different models and variables are consistently being selected by each technique. In addition, the persistence pattern for variables selected in overlapping and successive windows shows up more clearly in the models selected by the Bayesian technique. Once again this may indicate that the stepwise regression procedure is not selecting the 'best' variables.

Tables 7 and 8 contain the results showing the in-sample accuracy for both methods of estimation over each of the five year windows considered. These results are again very comparable. When Pr was greater than 0.6 or less than 0.4, on average the Bayesian technique accurately predicted earnings increase 72 percent of the time. The corresponding percentage for stepwise regression is 70.

Tables 9 and 10 show the forecast accuracy for each technique. The results are very similar, in fact when Pr was greater than 0.6 or less than 0.4 both methods accurately forecasted earnings 61 percent of the time.

4.3. Australian companies. The results for the Australian companies are summarised in tables 11-15. Table 11 presents the variables selected by each method. Very different models are again being selected by each technique. Also the persistence pattern for variables selected in overlapping windows shows up more clearly in the models selected by the Bayesian technique.

Tables 12 and 13 contain the results showing the in-sample accuracy of both methods of estimation over each of the five year windows considered. These results are again very comparable. However, when Pr was greater than 0.6 or less than 0.4, on average the Bayesian technique accurately predicted earnings increase 77 percent of the time. The corresponding percentage for stepwise regression is 80.

Tables 14 and 15 show the forecast accuracies for each technique. The results are very similar. When Pr was greater than 0.6 or less than 0.4 the MCMC method accurately forecasted earnings 68 percent of the time. The corresponding percentage for stepwise regression is 67.

5. SUMMARY

This paper presents a Bayesian MCMC sampling scheme that performs variable selection in logistic regression. The method is applied to the model used by Ou and Penman (1989), who employed stepwise regression to select variables from financial statements in order to predict and forecast company earnings direction. For US, UK and Australian companies with relevant information available on COMPUSTAT, stepwise regression is compared to the MCMC method presented here. The results are very similar for all three countries, with slightly better overall forecast performance for the Bayesian technique. The models and variables selected by the Bayesian method also appear to be more consistent from year to year, especially in overlapping sample periods, perhaps illustrating the deficiencies of the stepwise regression technique.

REFERENCES

- Gerlach, R. and Kohn, R.: 1998, A comparison of three Bayesian estimators of non-Gaussian state space models, manuscript being prepared for submission.
- Mira, A. and Tierney, L.: 1997, On the use of auxilliary variables, manuscript being prepared for submission.
- Ou, J. A. and Penman, S. H.: 1989, Financial Statement Analysis and the Prediction of Stock Returns, *Journal of Accounting and Economics* **11**, 295–329.
- Smith, M. and Kohn, R.: 1996, Nonparametric regression using Bayesian variable selection, *Journal of Econometrics* **75**, 317–343.
- Weisburg, S.: 1985, *Applied Linear Regression*, second edn, John Wiley and sons, inc., pp. 214–215.

Variable	Forecast Year									
	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
2 Δ sales	MS			M	M	MS	MS	M		
4 Δ Assets/inventory	M		S			M				M
5current ratio							S			
6 Δ above	M	M	MS	MS	MS			MS		
7inventory	S	M				M	MS			
8 Δ inventory		M	M	M	MS	MS	MS	MS	M	M
9invt. turnover		S	S					M		M
10 Δ above							S			MS
11quick ratio							S		M	
12 Δ quick ratio	M	S	S			MS	MS		MS	M
13sales/assets	M	M	M			M			MS	M
14 Δ above	S	S		M	M	MS	MS	M	M	
15return on assets	M	MS	M	M	M	M	MS	M	M	MS
16 days sales in AR	S		S				M	M		
17 Δ above	M	S	M	M		S	S			
18return on equity	M	M	MS				S		M	M
19 Δ above		S								
20 Δ cap.exp./assets	M		MS	M	M	MS	MS	MS	MS	M
21above lagged one yr.	M	S	M	MS	MS	M	MS	MS	MS	MS
23 Δ debt/equity				M	M	MS	MS		M	
24LT debt/equity		S								
25 Δ above		M	M				S	MS	MS	MS
26equity/fixed assets						S	S			
27 Δ equ./ fixed assets				M		MS	MS		S	M
28times int. covered							S			
29 Δ above					S	MS	MS	M	M	M
30ret. closing equ.	MS	MS	M	MS			S	M	MS	MS
31gross margin			S				S			M
32 Δ above							S			MS
33op. profit/sales			M	M	M	M	MS	MS	MS	MS
35pre tax income		M	M	M		MS	M	M	MS	M
36 Δ above					S		S			S
37net profit margin	MS	MS	MS	M	MS	MS	S	S	S	
38 Δ above		MS					S			
39sales/cash							S			
40sales/AR						M	M			
41sales/invt.		S	M	M				M	S	M
42 Δ above	M	MS	M	M	M	M	MS	M	MS	M
44 Δ sales/working cap.	M	S	M				S			
45sales /fixed assets						M	M	M	M	
46 Δ assets	MS	M	M	M	M	MS	MS	M	M	MS
47cash flow to debt					S		S			
48working capitol/assets	M	M	M		M			M		
50oper. inc./assets				S	M	M	M	M		
53issuance/tot. LT debt					M	M				
55cash div./cash flow					M		S			
56 Δ work capitol					S	S	S			
57net inc./cash flow	M	MS	M	M	MS				S	MS

TABLE 1. Variables selected by each estimation technique for the US COMPUSTAT database. An 'M' indicates that this variable was selected by the MCMC variable selection method and 'S' indicates it was selected by the stepwise regression method.

Period	Pr > 0.8	Pr > 0.6	Pr > 0.5	\bar{y}	Pr < 0.5	Pr < 0.4	Pr < 0.2	$1 - \bar{y}$
1992-96	88	72	62	56	58	62	80	44
1991-95	86	73	61	55	56	64	68	45
1990-94	83	73	63	51	59	66	79	49
1989-93	80	73	64	48	60	67	82	52
1988-92	78	73	64	49	59	65	87	51
1987-91	77	74	63	50	58	69	81	50
1986-90	83	75	63	52	58	67	82	48
1985-89	85	78	59	52	58	64	82	48
1984-88	85	73	61	55	56	66	79	45
1983-87	84	75	62	53	59	67	83	47
Avg.	83	74	62	52	58	66	80	48

TABLE 2. In sample accuracy for US companies using the MCMC variable selection technique. Pr is the estimated probability of an earnings increase ($y=1$). The numbers are percentage of time that the model accurately predicted the observations.

Period	Pr > 0.8	Pr > 0.6	Pr > 0.5	\bar{y}	Pr < 0.5	Pr < 0.4	Pr < 0.2	$1 - \bar{y}$
1992-96	79	70	58	56	57	61	72	44
1991-95	86	72	61	55	57	62	73	45
1990-94	79	71	64	51	58	63	78	49
1989-93	80	73	65	48	61	65	81	52
1988-92	68	75	63	49	58	67	100	51
1987-91	76	76	62	50	58	65	73	50
1986-90	78	73	59	52	56	64	77	48
1985-89	79	75	61	52	58	63	72	48
1984-88	82	74	60	55	57	72	100	45
1983-87	85	74	63	53	58	69	82	47
Avg.	79	73	62	52	58	64	78	48

TABLE 3. In sample accuracy for US companies using the stepwise regression variable selection technique. Pr is the estimated probability of an earnings increase ($y=1$). The numbers are percentage of time that the model accurately predicted the observations.

Period	Pr > 0.8	Pr > 0.6	Pr > 0.5	\bar{y}	Pr < 0.5	Pr < 0.4	Pr < 0.2	$1 - \bar{y}$
1997	57	55	53	47	54	54	53	53
1996	89	75	65	48	58	71	80	52
1995	87	68	63	58	55	55	50	42
1994	84	81	74	64	40	46	56	36
1993	82	81	72	59	51	62	75	41
1992	77	74	65	52	60	66	67	48
1991	68	54	50	40	70	80	57	60
1990	83	68	49	42	68	68	75	58
1989	79	70	56	48	61	71	62	52
1988	80	73	62	53	56	59	63	47
Avg.	79	70	61	51	57	63	64	49

TABLE 4. Forecast accuracy for US companies using the MCMC variable selection technique. Pr is the estimated probability of an earnings increase ($y=1$). The numbers are percentage of time that the model accurately predicted the observations.

Period	Pr > 0.8	Pr > 0.6	Pr > 0.5	\bar{y}	Pr < 0.5	Pr < 0.4	Pr < 0.2	$1 - \bar{y}$
1997	75	60	46	47	44	56	61	53
1996	89	72	57	48	62	72	100	52
1995	82	75	69	58	47	50	50	42
1994	87	81	74	64	41	44	50	36
1993	67	81	73	58	49	57	100	42
1992	72	73	64	52	59	61	60	48
1991	52	66	48	40	69	67	63	60
1990	73	64	51	42	67	67	56	58
1989	44	49	48	48	71	50	0	52
1988	81	72	63	53	55	60	62	47
Avg.	75	69	59	51	56	61	58	49

TABLE 5. Forecast accuracy for US companies using the stepwise regression variable selection technique. Pr is the estimated probability of an earnings increase ($y=1$). The numbers are percentage of time that the model accurately predicted the observations.

Variable	Forecast Year					
	1993	1994	1995	1996	1997	1998
2 Δ sales	MS	M	M	MS		M
4 Δ Assets/inventory						M
6 Δ current ratio						M
8 Δ inventory		M				S
13 Δ sales/assets	S	MS	M			M
14return on assets	M	MS	MS	MS	MS	MS
16debt/equity			S	S		
18above, long term					S	
21 Δ equ./ fixed assets	S	S		S	MS	MS
22 times int. covered		M				
23 Δ above				S		M
24 ret. closing equ.	MS			M		S
26 Δ gross margin						S
27op. profit/sales						MS
28 Δ above						MS
29pre tax income	M	M	M			M
30 Δ above	S		S			
31net profit margin	S	S		S		MS
32 Δ profit margin						S
35 Δ sales/invt		S				
39 Δ assets	MS	MS	MS	M		
43oper. inc./assets		MS	M	M	M	MS
44 Δ above	M					M
47 Δ deprec.	MS	MS	MS	MS	MS	MS

TABLE 6. Variables selected by both estimation techniques for the UK companies in the Global Vantage database. An 'M' indicates that this variable was selected by the MCMC variable selection method and 'S' indicates it was selected by the stepwise regression method.

Period	Pr > 0.8	Pr > 0.6	Pr > 0.5	\bar{y}	Pr < 0.5	Pr < 0.4	Pr < 0.2	$1 - \bar{y}$
1993-97	88	74	67	62	62	71	71	38
1992-96	84	72	63	60	57	67	60	40
1991-95	85	73	66	56	64	64	78	44
1991-94	87	75	69	55	66	71	78	45
1991-93	90	75	68	48	68	72	81	52
1991-92	70	70	62	43	66	72	84	57
Average	87	73	66	54	64	71	80	46

TABLE 7. In sample accuracy for UK companies using the MCMC variable selection technique. Pr is the estimated probability of an earnings increase ($y=1$). The numbers are percentage of time that the model accurately predicted the observations.

Period	Pr > 0.8	Pr > 0.6	Pr > 0.5	\bar{y}	Pr < 0.5	Pr < 0.4	Pr < 0.2	$1 - \bar{y}$
1993-97	90	75	66	62	60	66	69	38
1992-96	83	72	64	60	41	38	33	40
1991-95	84	74	64	56	61	65	86	44
1991-94	88	75	69	55	66	70	81	45
1991-93	93	77	68	48	68	72	75	52
1991-92	73	75	67	43	67	73	86	57
Average	88	75	66	54	61	64	78	46

TABLE 8. In sample accuracy for UK companies using the stepwise regression variable selection technique. Pr is the estimated probability of an earnings increase ($y=1$). The numbers are percentage of time that the model accurately predicted the observations.

Period	Pr > 0.8	Pr > 0.6	Pr > 0.5	\bar{y}	Pr < 0.5	Pr < 0.4	Pr < 0.2	$1 - \bar{y}$
1998	67	56	47	48	52	65	62	52
1997	86	68	61	57	58	62	100	43
1996	82	56	55	50	57	58	0	50
1995	88	72	71	69	34	33	42	31
1994	87	85	81	75	34	36	62	25
1993	91	79	72	60	63	65	54	40
Average	85	74	68	61	45	47	55	39

TABLE 9. Forecast accuracy for UK companies using the MCMC variable selection technique. Pr is the estimated probability of an earnings increase ($y=1$). The numbers are percentage of time that the model accurately forecasted the observations.

Period	Pr > 0.8	Pr > 0.6	Pr > 0.5	\bar{y}	Pr < 0.5	Pr < 0.4	Pr < 0.2	$1 - \bar{y}$
1998	70	60	53	48	67	69	75	52
1997	85	65	58	56	49	49	43	44
1996	87	68	55	52	54	53	25	48
1995	89	72	72	69	34	32	40	31
1994	81	86	83	75	30	32	44	25
1993	92	78	70	60	63	59	55	40
Average	86	72	65	60	50	49	45	40

TABLE 10. Forecast accuracy for UK companies using the stepwise regression variable selection technique. Pr is the estimated probability of an earnings increase ($y=1$). The numbers are percentage of time that the model accurately forecasted the observations.

Variable	Forecast Year					
	1993	1994	1995	1996	1997	1998
2 Δ sales	S		S			
3Inventory/assets						M
4 Δ above				MS		
8 Δ inventory		S	S	M		
9quick ratio						S
11sales/assets						M
12 Δ above		S	S			M
13return on assets	MS	MS	MS	MS	MS	MS
18long term debt/equity	S	MS	MS	S		
20 Δ equ./ fixed assets			S	S		
23 ret. opening equ.	M	MS		S		S
24op. profit/sales					S	M
25 Δ above		MS				
26pre tax income/sales		M	MS	MS	MS	M
27 Δ above	M			M		
28net profit margin				M	MS	M
33 Δ sales/work. capitol			M			
36 Δ assets				M		
37cash flow/debt				M	MS	MS
39 Δ work capitol/assets			M			
40oper. inc./assets					S	MS
43 Δ work. capitol			M	M	S	

TABLE 11. Variables selected by both estimation techniques for the Australian companies in the Global Vantage database. An 'M' indicates that this variable was selected by the MCMC variable selection method and 'S' indicates it was selected by the stepwise regression method.

Period	Pr > 0.8	Pr > 0.6	Pr > 0.5	\bar{y}	Pr < 0.5	Pr < 0.4	Pr < 0.2	$1 - \bar{y}$
1993-97	83	78	65	51	64	72	94	49
1992-96	83	81	65	55	61	76	90	45
1991-95	91	76	69	55	63	76	100	45
1991-94	88	77	70	56	70	73	92	44
1991-93	86	82	73	51	68	73	100	49
1991-92	94	88	80	49	72	80	100	51
Avg.	87	79	68	53	65	75	97	47

TABLE 12. In sample accuracy for Australian companies using the MCMC variable selection technique. Pr is the estimated probability of an earnings increase ($y=1$). The numbers are percentage of time that the model accurately predicted the observations.

Period	Pr > 0.8	Pr > 0.6	Pr > 0.5	\bar{y}	Pr < 0.5	Pr < 0.4	Pr < 0.2	$1 - \bar{y}$
1993-97	81	80	65	51	66	71	83	49
1992-96	80	84	66	55	62	75	100	45
1991-95	84	80	69	55	64	70	88	45
1991-94	86	77	70	56	65	76	94	44
1991-93	83	81	71	51	67	88	100	49
1991-92	94	88	78	49	68	87	93	51
Avg.	84	82	70	53	65	78	92	47

TABLE 13. In sample accuracy for Australian companies using the stepwise regression variable selection technique. Pr is the estimated probability of an earnings increase ($y=1$). The numbers are percentage of time that the model accurately predicted the observations.

Period	Pr > 0.8	Pr > 0.6	Pr > 0.5	\bar{y}	Pr < 0.5	Pr < 0.4	Pr < 0.2	$1 - \bar{y}$
1998	100	92	61	41	86	90	100	59
1997	67	61	47	37	76	92	100	63
1996	69	71	56	44	70	81	67	56
1995	80	63	58	52	56	50	50	48
1994	100	86	79	71	35	36	33	29
1993	68	64	62	55	53	61	83	45
Avg.	78	71	59	50	59	64	60	50

TABLE 14. Forecast accuracy for Australian companies using the MCMC variable selection technique. Pr is the estimated probability of an earnings increase ($y=1$). The numbers are percentage of time that the model accurately predicted the observations.

Period	Pr > 0.8	Pr > 0.6	Pr > 0.5	\bar{y}	Pr < 0.5	Pr < 0.4	Pr < 0.2	$1 - \bar{y}$
1998	100	87	55	44	83	90	100	56
1997	57	60	51	37	82	87	100	63
1996	80	72	57	41	72	80	100	59
1995	71	73	60	52	57	61	75	48
1994	100	79	73	71	31	24	50	29
1993	70	68	70	56	51	52	49	44
Avg.	79	73	61	50	63	60	54	50

TABLE 15. Forecast accuracy for Australian companies using the stepwise regression variable selection technique. Pr is the estimated probability of an earnings increase ($y=1$). The numbers are percentage of time that the model accurately predicted the observations.