Inertia and Discreteness:
Issues in Modeling Analyst Coverage

Maureen F. McNichols
Graduate School of Business
Stanford University
Stanford, CA 94305
(650) 723-0833
fmcnich@leland.stanford.edu

Patricia C. O'Brien
School of Accountancy
University of Waterloo
200 University Ave. West
Waterloo, Ontario
Canada N2L 3G1
(519) 888-4567 x5423
pobrien@uwaterloo.ca

Comments welcome

September 2001

The authors gratefully acknowledge the financial support of the Stanford GSB Financial Research Initiative, London Business School, York University, the University of Waterloo, the University of Amsterdam, the Canadian Academic Accounting Association, and the Canadian Social Sciences and Humanities Research Council. We thank David Stolin, Donglei Chen and Jerry Sun for research assistance. We are grateful to *I/B/E/S International* for the use of *I/B/E/S* analyst forecast data, and to Don Cram for setting up the data and sharing his *SAS* macros. We have benefited from comments on an earlier paper, "Analyst Coverage, Firm Performance and Earnings Forecast Errors," by seminar participants at the 1997 Stanford Accounting summer camp, at the Universities of British Columbia and Washington, and at York University; and from comments on this paper from the referee, as well as from seminar participants at Duke University, the University of Toronto, the University of Waterloo, and the 2001 AAA annual meetings.

1. Introduction

Merton's (1987) theory of information dispersal suggests that if investors rely on brokers to learn about their investment options, then firms with wider coverage will be more valuable, because a larger investor base knows of the investment opportunity. Several studies provide support for Merton's theory, demonstrating that firms with more analyst coverage have lower trading costs[1] and more trading volume,[2] and that firms' disclosure policies reduce their financing costs only if analyst coverage is low.[3] Bhushan (1989), in a related vein, proposed analyst coverage as a proxy for economy-wide investment in information about a firm. Since then, coverage or following has come into widespread use as an indicator of investor information. This in turn has led to interest in analysts' motives and decision processes. Subsequent studies seeking to explain how analyst coverage varies across firms explore a rich set of motivating factors, such as fees from institutional investors,[4] the quality of firm-level disclosure,[5] firm performance[6] and investments in intangible assets.[7]

In this paper we explore the effects of two specification issues in empirical models of analyst coverage. First, because analysts tend to cover and ignore the same firms consistently year after year, we examine the effects of this inertia on models of coverage. Second, we examine the use of the standard linear regression model, which assumes the dependent variable is continuous. Analyst following is categorical and has a highly skewed distribution bounded below by zero, so the assumption of a continuous distribution is tenuous. We explore alternative models designed for discrete data, namely ordered probit and logit, to see how closely linear regression models compare with these alternatives. We are motivated to study these two

---

[1] Brennan and Subrahmanyam (1995)
[2] Alford and Berger (1999)
[3] Botosan (1997)
[4] O'Brien and Bhushan (1990)
[5] Lang and Lundholm (1996)
[6] McNichols and O'Brien (1997)
[7] Barth, Kasznik and McNichols (1999)

specification issues because they reflect significant departures from the behavior of analysts assumed in the existing literature. Our purpose is to examine their effects on descriptive and predictive inferences drawn from this literature.

We find strong evidence that analyst inertia affects inferences in both time-series and cross-sectional studies of analyst following, because lagged following is a significant omitted variable. We find that both descriptive validity and predictive power of models of analyst coverage improve dramatically when we control for lagged analyst following.

We find that we can firmly reject the linear regression model in favor of discrete response models using the Vuong test statistic, and that descriptive inferences regarding volatility, market share and industry growth are affected by the model choice. The choice of model, however, does not materially affect either predictive accuracy or descriptive inferences about the effects of size, trading volume or share price on analyst coverage.

We explore these issues in a five-industry sample of firms, though the issues we raise about the distribution and persistence of analyst following are not unique to these industries. Primary criteria in selecting our sample industries are that they have substantial numbers of firms, moderately broad interest from the analyst community, and contain firms in a wide range of sizes. This allows for substantial variation in firm-level characteristics, and sufficient power to draw meaningful conclusions. Our industries are: Food, Plastics, Computers, Lab Instruments, and Pharmaceutical/Biotech.

We begin by constructing an empirical model of analyst following. We draw on prior research with models of analyst following, which we discuss in section 2. In section 3, we describe the two potential sources of mis-specification, and our methods of investigating them. We describe our data sources and sample in section 4, and provide results in section 5. Section 6 concludes.

2. *Literature on analyst following*

In one of the earliest papers to model analyst following, Bhushan (1989) conjectures that analysts will decide to cover firms by weighing the costs of effort expended in information gathering against the benefits of brokerage commissions. He finds that analyst following as reported in *Nelson's Directory* is positively related to the number of institutions holding the firm's shares, the percent of the firm held by institutions, its return variability and its size; and is negatively related to the firm's number of lines of business.

Extending this idea, O'Brien and Bhushan (1990) examine whether these correlations reflect causal relations. They use a simultaneous equations model to study the interaction of analysts' and institutions' decisions, and find that changes in analyst following are positively related to net entry into the industry, negatively related to pre-existing analyst following, and higher for regulated industries. The effects of institutional ownership and firm size are not distinguishable from each other, once the simultaneous relation between analysts and institutions is taken into account. The stock's residual variance and the lagged number of institutions do not affect following.

Brennan and Hughes (1991) conjecture that analysts are attracted to low-priced stocks because brokerage commissions are proportionally higher for these stocks. They find that *I/B/E/S* analyst following increases with size and return variance, and decreases with price, and with current and lagged stock returns.

Rajan and Servaes (1997) employ a two-stage censored-data model to control for lack of *I/B/E/S* coverage. Their first-stage model of the determinants of *I/B/E/S* coverage is relevant for our study. They find that for IPO firms, inclusion in *I/B/E/S* is related to firm size, trading on a major exchange (*NYSE*, *AMEX* or *NASDAQ*), and the number of firms in the industry. They examine other variables (trading volume, turnover, stock price, volatility and industry growth), but only report that their second-stage results on IPO characteristics are insensitive to the inclusion of these variables, not whether the variables themselves are significant determinants in the first stage.

Alford and Berger (1999) model analyst following, forecast accuracy and trading volume as simultaneous determinants of firms' information environments. They find that analyst following is positively associated with accuracy and trading volume and higher for regulated industries. They also find that volume is not endogenous to analyst following, though accuracy is.

Rock, Sedo and Willenborg (2000) replicate Bhushan (1989)'s study, using alternative assumptions about the error distribution based on arrival-time models. That is, they model counts of analyst following as counts of incidents whose arrival rates have a Poisson or negative binomial distribution. They find that the arrival time models exhibit a better fit to the data than the regression model employed by Bhushan, and that inferences about the number of institutions holding shares are affected by the choice of model.

All of the studies cited above, with the exception of O'Brien and Bhushan (1990) who use a first-differenced specification, model analyst following using the level of following as the dependent variable, without considering the potential effect of lagged following. All of the studies cited, with the exception of Rock, Sedo and Willenborg (2000) who use arrival-time models appropriate for count data, treat analyst following as if it were a continuous variable.

3. Specification issues in empirical models of analyst coverage

Our goal is to demonstrate the effects of two specification issues in empirical models of analyst following. First, analyst following exhibits inertia: it tends to be highly autocorrelated. Second, analyst following is categorical and has a highly skewed distribution. We discuss each of these issues in turn below.

3.1 Analyst inertia

Analysts could conceivably collect and produce information for entirely new and separate firms each year. In practice, of course, the costs and benefits of engaging in information collection or production will likely depend on company traits that are reasonably stable through time. For example, information is more accessible from a larger number of sources (reducing

analysts' costs) for larger firms than for smaller firms.[8] Brokerage commissions to the analyst's employer, often cited as a factor in analyst coverage decisions,[9] increase with trading volume, which is generally higher for larger firms. Firms can choose disclosure policies that foster or hinder analyst access,[10] and these policy choices remain relatively stable from year to year. Industry-specialist analysts may find it necessary to report consistently on the major firms in the industry to retain credibility in their specialty.

This explanation for inertia, stability in the underlying traits that cause following, suggests that the misspecification is serial correlation in the level of analyst following. To the extent that this stability is not perfectly captured by the levels of the independent variables included in the model, the regression model errors will remain serially correlated, and the regression model will be misspecified.

This phenomenon is of course not unique to the analyst setting that we examine. An extensive literature in macroeconomics addresses the serial correlation in various economic series, such as GNP, consumption and price levels. Examples include Nelson and Plosser (1982), Stock and Watson (1986) and Schwert (1987).[11] An extensive literature in financial economics examines the serial correlation in stock prices and returns. Examples include Lo and MacKinlay (1988), Fama and French (1988) and Poterba and Summers (1988).[12] Relatedly, a large literature in econometrics explores the issues associated with serial correlation in the dependent variable and its implications for model specification and hypothesis testing.[13]

Although the phenomenon of serially correlated levels is well known to financial economists, the literature on analyst following published in major journals from 1989 through 2001 has largely ignored the issue. Bhushan (1989), Brennan and Hughes (1991), Rajan and

[8] See Bhushan (1989).
[9] See, e.g. Bhushan (1989), Alford and Berger (1999).
[10] See Lang and Lundholm (1996).
[11] See Diebold and Nerlove (1990) and Campbell and Perron (1991) for overviews of this literature.
[12] See Campbell, Lo and MacKinlay (1997) ch. 2 for an overview of this literature.
[13] See Greene (1997) for an overview.

Servaes (1997), Alford and Berger (1999), Rock, Sedo and Willenborg (2000) all model analyst coverage in level form, without considering the potential effect of the lagged level of following.

Aside from stability in underlying causal factors, structural characteristics of the analyst's job undoubtedly also play a role in inducing inertia in coverage. Analysts face start-up costs, such as costs of learning about the firm's products, processes and sources of competitive advantage.[14] Access to in-depth information may depend upon the analyst's spending time and effort developing and maintaining relationships with company employees. To the extent that these costs play a role in analysts' decisions, an analyst is more likely to retain coverage of a firm he or she has covered in the past than to initiate coverage of a new firm, all else equal. This additional rationale for inertia as a distinct causal factor suggests that it is an omitted economic variable, and not merely an accumulation of imperfectly modeled stability in other causal factors. The stability in other underlying causal factors, however, implies that inertia will be correlated with other factors in the model, potentially impairing inferences.

The result of environmental stability and structural aspects of analysts' work is that, empirically, coverage is "sticky." This is evident in our data, where the lagged level of analyst following has a 96% Pearson (89% Spearman) correlation with current-year following, stronger than any other univariate relation we measure.[15]

We model this mis-specification as an omitted variable in a levels specification, rather than as serial correlation, for two reasons. First, the levels specification enhances comparability with the earlier work that motivates this study. It allows us to demonstrate the misspecification by testing whether the other regressors may adjust for the effect of lagged following, and to explore the effects on inferences. If inertia is solely due to stability in other causal factors, and the included factors measured in levels control for this stability, then lagged following will add nothing new to the model. Second, we wish to stress the fact that omitting lagged following affects cross-sectional models as well as those that pool time-series and cross-section data.

[14] McNichols and O'Brien (1997) and Hayes (1998) discuss start-up costs of initiating coverage. Mikhail, Walthers and Willis (1997) discuss analyst learning.
[15] We describe our data in section 4.

Alternatives to our omitted variable specification include estimating the model in differenced form, or using a transformation procedure like the Cochrane-Orcutt (1949) transformation. Both adjust the variables for serial correlation prior to estimating the regression. The former assumes that the auto-regressive coefficient is 1.0, which runs the risk of over-differencing if the true auto-regressive coefficient is less than 1.0.[16] As we describe in section 5, though the auto-regressive coefficient is 0.96, it is reliably less than 1.0. Less restrictive transformation procedures estimate the auto-regressive serial correlation separately from the other parameters. Though they allow for coefficients different from 1.0, they do not allow us to test the extent to which other factors used in past studies may absorb the effect of lagged following.

Generally, omitted variables result in biased and inconsistent estimators of both coefficients and standard errors of correlated regressors. In this case, since the omitted variable is the lagged dependent variable, the coefficient estimates will still be biased in finite samples after we include lagged following, but the estimates will be consistent.[17]

We demonstrate the importance of analyst inertia by estimating models of analyst coverage with and without lagged following. We draw on prior research for most of our exogenous variables: firm size, exchange listing, trading volume, volatility, price and industry growth. We also include indicators for the industry and year, and a market share variable to capture the firm's relative importance in the industry. Rajan and Servaes (1997) find that analyst following varies with year and exchange listing. Brennan and Hughes (1991) and Alford and Berger (1999) hypothesize that analysts seek to generate brokerage commissions, and can more readily do so for high-volume stocks. Bhushan's (1989) rationale for including volatility is that information is more valuable for volatile stocks. Brennan and Hughes (1991) conjecture that analysts prefer low-priced stocks. Following O'Brien and Bhushan (1990), we include industry

[16] See O'Brien and Bhushan (1990) for a model of analyst following using first-differenced data.

[17] Including the lagged dependent variable as a regressor induces coefficient bias because the lagged dependent variable y_{it-1} cannot be uncorrelated with the error vector $\underline{\varepsilon}$. In particular, y_{it-1} cannot be uncorrelated with ε_{it-1} or any prior $\varepsilon_{it-\tau}$, $\tau>0$. The OLS estimate is consistent because y_{it-1} is uncorrelated with the current and future $\varepsilon_{it+\tau}$, $\tau\geq 0$. See e.g. Kennedy (1992) p. 141, and Schmidt (1976) p. 40.

growth, measured as the five-year industry growth in sales, to capture cross-industry differences in coverage levels. Though most prior studies include size in their models of analyst coverage, we find the economic rationale for including size less than compelling. We follow the literature in including size, but also define a related variable whose rationale we find more plausible: market share is the firm's sales, as a proportion of industry sales for the year. The rationale for market share is that industry-specialist analysts may find it necessary to issue reports on the industry's dominant firms to maintain their credibility as specialists.

We estimate the following model of the number of analysts following firm *i* in year *t*:

$$\begin{aligned} Analysts_{it} = a_{0t} + a_{1industry} + a_{2exchange} + a_3 Size_{it} + a_4 Volume_{it} + a_5 Volatility_{it} \\ + a_6 MktShare_{it} + a_7 Price_{it} + a_8 IndGrowth_{it} + [a_9 Analysts_{it-1}] + \varepsilon_{1it} \end{aligned} \quad \textbf{(1)}$$

We describe our data sources and variable definitions in section 4.

The omitted variable hypothesis is straightforward to test, because the models including and excluding the variable are nested. Under the null hypothesis that lagged following is not an omitted variable, we test $\alpha_9 = 0$ in equation (1). We also show the relative importance of lagged analysts by comparing equation (1) with the much simpler model that analyst coverage is a linear function only of lagged following:

$$Analysts_{it} = a_0 + a_9 Analysts_{it-1} + \varepsilon_{2it} \quad \textbf{(2)}$$

This simpler model has the advantage of less onerous data requirements, though obviously it lacks richness.

In addition to using measures of statistical fit (adj. R^2, t- and F-statistics), which may be compromised by model mis-specification, we compare the different models on predictive accuracy in a holdout sample. This allows us to assess the economic relevance of model differences, and to escape the specification issues of particular models. We describe how we construct and compare the holdout sample predictions in section 3.3.

We demonstrate that omitting lagged analyst following is a critical mis-specification in this context. When we include lagged following along with variables shown by earlier research

to explain analyst following, we find that lagged following dominates all other regressors in terms of its statistical significance, and that predictive accuracy in a hold-out sample improves markedly.

3.2 Discrete and skewed distribution

The second specification issue we explore is the highly skewed and discrete nature of analyst following data, which by definition must be non-negative integers. In practice, analyst following ranges in value from 0 to about 50 for a given firm-year. Most prior research has treated the data as if they were continuous, not discrete, presumably on the assumption that the range was broad enough to approximate a continuous distribution. The difficulty with this assumption is that the distribution is also highly skewed toward low values, so that the effective data distribution for most firms covers only a few values. In our data, 40.3% of firm-years have no coverage listed in *I/B/E/S*. The modal level of coverage for covered firms is 1 analyst (12.1% of firm-years), and the number of observations declines rapidly as the number of analysts increases.

Rock, Sedo and Willenborg (2000) model analyst following using Poisson and negative binomial regression. That is, they model counts of analyst following as counts of incidents whose arrival rates have a Poisson or negative binomial distribution. We do not find these models appealing for analyst data. In particular, the models require the assumption that counts in non-overlapping time periods are independent, which as we point out in our discussion of inertia in section 3.1, does not characterize analyst coverage decisions. The implication of this assumption is that the arrival-time models are not appropriate for panel data.[18] Rock, Sedo and Willenborg (2000) limit their study to replicating Bhushan (1989), who used a cross-section at a single point in time, so they do not confront the issue of multiple time periods. The literature subsequent to Bhushan (1989) uses panels of several years, so the arrival-time models' assumption of intertemporal independence limits their wider applicability.

[18] See Rock, Sedo and Willenborg, p. 356.

We address the issue of discreteness by estimating our model of analyst following using ordered probit and logit, which allow us to specify the dependent variable simply as a set of ordered categories. These methods do not require symmetry or uniform numbers of observations per category, so skewness is an issue only insofar as some categories may have too few observations for estimation. For higher levels of analyst following where observations are sparse, we combine adjacent levels to ensure enough observations per category to allow estimation.[19]

In contrast to the standard regression model's distributional assumption on the errors from the linear model, ordered probit or logit imposes an assumption on the unmodeled portion of the probability that determines the distribution of observations across categories. The probit model assumes a cumulative normal probability distribution, while the logit model assumes a cumulative logistic probability distribution.[20] The only requirements on the data are that the dependent variable categories be ordered, as indeed counts of analyst following are.

Because the ordered probit and linear regression models employ quite different distributional assumptions, the models are not nested. We use the test devised by Vuong (1989), and introduced in the accounting literature by Dechow (1994) to compare the linear regression with the probit models. The Vuong test statisic firmly rejects linear regression in favor of probit.

To address the issue of the economic significance of model difference, we perform out-of-sample predictions from the various models, using two different loss functions, to compare the models. We describe these comparisons in section 3.3.

[19] Several readers of earlier drafts of this paper proposed Tobit regression to deal with the fact that analyst following is non-negative. The Tobit model, however, assumes a continuous distribution, so we did not pursue this alternative.
[20] Logit and probit give substantially identical results, so we report only probit results in this paper. Logit results are available from the authors on request. Hereafter, we refer only to probit.

3.3 Out-of-sample prediction errors and comparisons

To compare the different models' predictive ability outside the estimation sample, we randomly assign 25% of firm-year observations to a hold-out sample, and use the remaining 75% for estimation. We impose the same data requirements on all the models, so that the estimation and prediction samples for our comparisons are identical.

We construct predictions from the linear regression model in the standard way, by applying the estimated coefficients to the holdout sample data. For example, we construct predictions from equation (1) as follows:

$$Prediction_{it}^{reg} = \hat{a}_{0t} + \hat{a}_{1industry} + \hat{a}_{2exchange} + \hat{a}_3 Size_{it} + \hat{a}_4 Volume_{it} + \hat{a}_5 Volatility_{it} + \hat{a}_6 MktShare_{it} + \hat{a}_7 Price_{it} + \hat{a}_8 IndGrowth + \hat{a}_9 Analysts_{it-1} \quad \textbf{(3)}$$

In (3), the $\hat{\alpha}$ are coefficient estimates based on the estimation sample. We generate predictions by applying these estimates to hold-out sample data.

Constructing predictions from probit models is somewhat more complex. The probit models produce coefficient estimates and cut-off points for each category of the dependent variable. The first column of Table 1, Panel C shows our 15 categories of analyst following. We use the coefficient estimates to compute a score for each hold-out observation, and use the score and the estimated cut-off points to obtain predicted marginal probabilities in each category: $\hat{p}_{itk} = \Pr[Analysts_{it} \in k]$, k=1,15. We then compute the predicted level of analyst following for the observation by multiplying the marginal probability by the level of following in category k,[21] and summing over all categories:[22]

$$Prediction_{it}^{probit} = \sum_{k=1}^{15} \hat{p}_{itk} Value_k \quad \textbf{(4)}$$

[21] See Table 1, Panel C for our categories. Each of the first 6 categories, 0 through 5, has only one value of analyst following. For the remaining 9 categories, we use the midpoint of the range of values. For example, for the 7th category, $Analysts_{it} \in [6,10]$, we use the value 8.

[22] We explored an alternative way to construct predictions from ordered probit and logit models. We selected the category with the highest predicted marginal probability p_{itk}, and used the level of following in that category as the predicted level of following. In a pilot sample of pharmaceutical and biotechnology firms, we found that equation (4) yielded more accurate predictions than this alternative.

We create prediction errors by subtracting the predictions from equation (3) or (4) from the observed value of analyst following for that observation:

$$Error_j = Analysts_j - Prediction_j^m, \qquad m \in [reg, probit] \tag{5}$$

where the index *j* ranges over all firm-year observations.

We compare the prediction errors using three metrics, representing two different loss functions: mean squared prediction error (MSE) and its square root (RMSE), and the mean of absolute percentage errors (MA%E), defined as follows:

$$MSE = \frac{1}{N}\sum_{j=1}^{N} Error_j^2 \tag{6}$$

$$RMSE = \sqrt{MSE} \tag{7}$$

$$MA\%E = \frac{1}{N}\sum_{j=1}^{N} \frac{|\, Error_j \,|}{\left(Analysts_j + 1\right)} \tag{8}$$

The first two of these, MSE and RMSE, are often used for comparisons in Monte Carlo tests. Since MSE can be re-written as [variance + bias^2], it allows for a trade-off between bias and variance in the predictions. That is, a biased estimator may be preferable to an unbiased one if it is more precise (has lower variance). RMSE is essentially the same measure, but facilitates discussions of economic significance because it is denominated in the same units as the dependent variable, in this case the number of analysts. Both these measures rely on a squared error loss function, which is appropriate in situations where larger errors are costlier than smaller errors, and an error's cost is independent of the value being predicted. Arguably, however, a prediction error of 1 analyst may be less costly when the actual number of analysts is 30 than when the actual number of analysts is 1. MA%E addresses this concern, by weighting the errors by the actual value of analyst following. We add one to the denominator to avoid dividing by zero when firms have no following.

4. Data, Variable Definitions and Descriptive Statistics

We use Compustat SIC codes to identify firms in 5 industries: 2010-2099 (Food and Kindred Products), 3080-3089 (Plastics), 3570-3579 (Computers), 3820-3829 (Lab Instruments), and 2830-2839 plus 8734 (Pharmaceutical and Biotech firms). We chose these industries based on three criteria. First, we examined the value-relevance of earnings for these industries to ensure that earnings are an important factor in valuation. Since our measure of analyst following (discussed below) is the existence of an analyst earnings forecast on *I/B/E/S*, we wished to choose industries where that report would be meaningful. Second, we selected industries with a sizeable number of firms at the 2-digit level, to permit within-industry analysis. Third, we chose industries that have experienced substantial growth in the number of firms, to allow us to study analyst coverage decisions in a setting with greater choice than more mature industries.

From the Compustat list of firms, we use CRSP to identify all the cusip numbers associated with these firms, and to identify the years in which their stock traded. We use the expanded cusip list to obtain *I/B/E/S* data for as many of these firms as possible.

We define analyst following to be the number of analysts predicting current year earnings listed on the *I/B/E/S* Summary file in the fiscal year end month. We use the end of the fiscal year, rather than an earlier point, because this measure of analyst following generally increases over the course of the fiscal year as analysts publish their reports, and peaks near the end of the year. If a firm is not listed on *I/B/E/S* in the fiscal year end month, then we define the number of analysts to be zero. We include firms with no analyst following, since their lack of following is a function of the same analyst decisions about the costs and benefits of coverage as positive levels of following for covered firms. In our sample, 40.3% of firm-years have no analyst coverage listed on *I/B/E/S*. Since we do not eliminate firms because of missing data on *I/B/E/S*, the universe of firms for our study is firms in our industries, as identified by Compustat SIC codes, that have some trading volume on *CRSP* during the years 1987 through 1996.[23]

[23] Most prior studies construct their samples of firms conditional on having some analyst coverage. Bhushan's (1989) sample is conditioned on existence of an analyst listing in *Nelson's Directory*, while O'Brien and Bhushan

Based on the prior literature on analyst following and our conjecture about inertia, we use the following factors to model analyst following: year, exchange, industry, trading volume, volatility, price, size, market share, industry growth and lagged analyst following. We use Compustat to determine the fiscal year end date for each firm, and adopt Compustat's fiscal year convention to group observations by year. We then use *CRSP* data to determine exchange listing at the end of the fiscal year (*Exchange*), the natural log of average daily trading volume over the fiscal year (*Volume*), the standard deviation of returns over the fiscal year (*Volatility*), stock price at the end of the fiscal year (*Price*) and the natural log of *CRSP* market capitalization at the end of the fiscal year (*Size*).[24] We use Compustat sales data to define industry growth (*IndGrowth*) as the average across industry firms in the percentage change in sales over the immediately prior 5-year period, and market share (*MktShare*) as the firm's sales as a percent of industry sales for the year.

We align the end-of-year number of analysts with explanatory factors from the current fiscal year, on the assumption that by the fiscal year-end month, analysts have essentially all of the information about the firm's financial position and trading activity for the current fiscal year, and they use this information in their decisions about whether to report. This may not be strictly true in every instance. For example, if the stock price moves dramatically in the last week of the fiscal year, our size and price measures will inaccurately measure the size and price information available to many analysts. A similar issue exists for any alignment of the data, however, because analysts do not report simultaneously.

Table 1 Panel A shows the distribution of sample observations across industries, before and after we impose the data restrictions. From a sample with 6,901 firm-year observations, we lose roughly 13% when we impose the requirement that complete data be available, leaving us with 5,985 firm-years. We randomly assign approximately 75% of observations to the estimation

---

(1990) and Brennan and Hughes (1991) condition their samples on listing in the I/B/E/S database. In this respect, these studies exclude a large part of the population of firms. We have replicated our results using samples conditioned on at least one analyst, without altering our inferences. Results available upon request.

[24] We examined an alternative size definition, *CRSP* decile ranking at the end of the calendar year, without changing any substantive inferences.

sample, leaving approximately 25% in our hold-out sample.[25] Table 1 Panel B shows the distribution of estimation and hold-out observations across industries.

For the probit models of analysts following, we convert following into 15 categories, as shown in Table 1 Panel C. This panel illustrates the extreme skewness of analyst following data toward low values, in all industries.

In Table 2 we report the distributions of sample observations by year and exchange, and by year and size decile. The number of firms in the sample shows an upward trend over time, but this is not accompanied by perceptible changes in the size or exchange distribution of firms. Between 1987 and 1996, the distribution of observations across exchanges has varied somewhat, but overall no clear pattern is evident. Likewise, the size decile breakdowns show some year-to-year variation during our sample period, but no clear directional pattern.

In Table 3, we report the median values of our regressors for each level of analyst following in 1991, roughly in the middle of our sample period. Size, volume, market share, price and lagged following all tend to increase with analyst following, while volatility tends to decrease. These relations are expected from prior research, except perhaps for the increasing relation between price and following. Brennan and Hughes (1991) suggest higher prices should be associated with lower following. Table 4 shows the cross-sectional correlations of the various determinants of analyst following in 1991. As expected for levels data, the variables are highly correlated. Large firms tend to have high analyst following, high volume, low volatility, high market share and high prices.

5. Results

5.1 The effects of analyst inertia

Table 5 shows the effects of correcting for analyst inertia. The first set of three columns in this table contains regression results, while the probit results appear in the second set of three

[25] We use the SAS uniform random number generator to generate a random number in the interval [0,1] for each observation. If the random number is less than 0.75, we assign the observation to the estimation sample, and otherwise to the hold-out sample.

columns. Within each set, the column marked I corresponds to the simple model (2) with only lagged following, while the columns marked II and III correspond to equation (1), respectively excluding and including lagged following. In the following discussion, we focus on the results from the standard regression model for the sake of simplicity, but except where noted inferences apply as well to the probit model.

From Table 5 it is clear that analyst following is strongly autoregressive. With no other covariates in the regression (column I), the regression coefficient on lagged following is a strongly significant .96 (t-statistic = 253.5). Note however, that it is also reliably different from 1.0 (t-statistic = 10.8), indicating that first-differencing may over-correct for lagged following. When we include other regressors in the model (column III), the regression coefficient drops to .85, but is still highly significant (t-statistic = 141.0), strongly rejecting the null hypothesis that $\alpha_9 = 0$. The goodness-of-fit (adjusted R^2, model F or X^2) statistics shown in Panel C of Table 5 indicate that the models that include lagged following, even the simple one with no other variables, fit the data better than those that ignore analyst inertia.

In Panel D of Table 5 we present the results of our predictions in the holdout sample. When we include lagged following in the model (column III), we obtain a RMSE of 1.9 analysts, as compared with 4.3 analysts when we fail to account for inertia. Recalling that more than 50% of sample firm-years have 0 or 1 analyst, we believe that this represents an economically meaningful improvement in prediction. Percentage errors, which adjust for the level of actual following, show that accounting for analyst inertia decreases errors from 133.6% of following to 40.0% of following on average. Again, we believe this decline to be economically meaningful.

We find it interesting that inferences about many of the other factors are unchanged by including lagged following, in spite of apparently high pairwise correlations shown in Table 4. The exceptions are market share and industry growth. Both these results also vary with the estimation method, and therefore we discuss them in more detail below where we compare regression and probit estimation.

Price is statistically significantly related to following in all our specifications. Our positive coefficient stands in contrast to the results of Brennan and Hughes (1991), who find a negative relation between price and analyst following. Their regression model includes current and lagged stock returns as regressors. We suspect these exert a strong collinear effect on price in their regression, since we also find a positive univariate relationship (see Table 4).

Statistically speaking, model III is better than model I: the additional variables improve the descriptive value of the model. This is evident both from the slight increase in adjusted R^2 and in statistically significant t-statistics on size, volume and price in column III. Whether the difference is economically meaningful is another matter. The hold-out sample results in Panel D of Table 5 show little difference between the two. RMSE is about 2 analysts in either case, while the percentage error from the simple regression model, at 38.8% of actual, is very slightly lower than its counterpart from the richer model, at 40.0%. The data requirements of the richer model reduce the sample size, a cost in loss of generality that must be weighed against the modest improvement in the model's predictive power. Table 1 Panel A indicates that the sample reduction is between 10 and 16% in our sample.

Another indicator of economic significance is the amount of change in a regressor required to increase following by one analyst. Using the coefficients in regression model III, we find that a firm would need to increase the natural log of size by 4.0, increase the natural log of volume by 3.3, or increase in price by $50, on average, to gain a single new analyst. To put these amounts in perspective, in our estimation sample the interquartile ranges of log of size, log of volume, and price are, respectively, 2.6, 2.6 and $15. These indicate that a firm would need to change from the bottom quartile to the top quartile in any of these characteristics to increase coverage by a single new analyst. This is largely a cross-sectional measure, since the number of firms in our estimation sample is two orders of magnitude larger than the number of years. Perhaps a more relevant measure is the year-to-year change in these variables for a given firm. A year-to-year change of one analyst is at the 72nd percentile of our estimation sample distribution,

indicating that 28% of our estimation sample have increases of 1 or more analysts per year.[26] The median year-to-year changes in log of size, log of volume and price are, respectively, 0.09, 0.1 and $0.03, which are considerably smaller than the amounts estimated above, 4.0, 3.3, and $50, respectively, that would be required to increase coverage by one analyst. The estimated increase in log of size to achieve a single new analyst, 4.0, is above the maximum of year-to-year changes in our sample, indicating an improbable amount of growth in size is needed to affect coverage. The estimated increase in log of volume and price, 3.3 and $50, are above the 98th and 99th percentiles, respectively, of year-to-year changes in our sample, indicating that the changes in these variables required to affect coverage are highly unusual. We conclude that the economic significance of these factors individually is low, despite their strong individual t-statistics.

Our overall conclusions are: (1) omitting lagged following from models of analyst coverage results in misspecification; (2) for purposes of prediction, a simple autoregressive model is roughly as good as models incorporating levels of other causal factors along with lagged following; (3) for purposes of description, several causal factors suggested in the prior literature, namely size, volume and price, remain statistically significant after including lagged following; (4) the economic significance of the causal factors suggested in the prior literature is quite modest.

5.2 Linear regression versus ordered probit

Table 5 also contrasts ordered probit estimation, in the second set of three columns, with linear regression in the first set of three columns. For the probit estimation, we transform analyst following into the set of 15 categories shown in Panel C of Table 1.

Our hold-out sample predictions in Panel D of Table 5 indicate that overall across models I through III, the probit results are quite similar to the regression results. They support the importance of including analyst inertia (model II versus model I or III), and show only minor

[26] An additional 16% of our estimation sample have decreases of 1 or more analysts per year. We focus on the upper tail of the distribution in our comparisons.

improvements from adding regressors beyond lagged analyst following (model I versus model III).

The hold-out sample comparison of probit versus regression shows little difference between the two estimation methods, once we adjust for inertia (model I or III, regression versus probit). In terms of RMSE, both methods produce average errors of approximately 2 analysts. MA%E, which adjusts for the level of actual following, shows some improvement from using probit in the case of model III (40.0% versus 34.5% of actual), but not in the case of model I (38.8% versus 45.1% of actual). These differences are in our view of questionable economic significance, and clearly are not of the same magnitude as the improvements from incorporating analyst inertia.

We find that size, trading volume, price and lagged following are all significant determinants of analyst following, regardless of the estimation method. The contributions of market share and industry growth appear to depend on the specification of the model, both in terms of the estimation method and in terms of whether lagged following is included. Volatility is not significantly related to following when we estimate the relation using regression, but is strongly significant when we use probit. We conclude that, though the model specification alters inferences about market share, industry growth and volatility, those relating to size, trading volume, price and lagged analyst following are not materially affected by the choice of estimation method.

The holdout sample comparisons reported above between regression and probit estimations show little economic difference between the two. In Table 6, we report a statistical comparison using the Vuong (1989) test statistic. Because of the sparse numbers of observations at higher levels of analyst following, to estimate the probit model we must transform the dependent variable to categories as shown in the first column of Table 1, Panel C. This, however, means that the two models shown in Table 5 are not directly comparable using the Vuong likelihood statistic. To construct the Vuong statistic, we therefore re-estimate the regression model using the probit category values of analyst following for the dependent variable,

rather than the raw values. Note that this transformation does not alter values of following between 0 and 5 inclusive, which are the majority of values in the sample. For larger values of following, we replace the level of following with the value of the category's lower bound. The resulting regression results are qualitatively similar to those we report in Table 5.

Vuong (1989) shows that the limiting distribution of the test statistic reported in Table 6 is the standard normal. The Vuong test provides a statistical measure for comparing the log-likelihood values, which are uniformly higher (less negative) for the probit estimations than for the regression estimations. Negative values of the statistic favor the probit model. The test strongly rejects the regression model in favor of the ordered probit model, with z-statistics of -47.8, -65.2 and –46.2 for models I, II and III respectively. Note that the pseudo-R^2 values for the probit models are uniformly *lower* than the adjusted R^2 values for the regression models, a reminder that this familiar summary statistic is not comparable across different types of estimation method.

In summary, we find strong statistical support for the superiority of probit estimation over regression for analyst data. Whether the statistically measurable difference is economically meaningful is, however, in doubt. Using criteria of root mean squared prediction error and prediction error as a percent of actual following, we find little substantive difference between the two estimation methods.

5.2 Year-by-year regressions

As a final demonstration of our point that analyst inertia operates as an omitted variable in cross-sections, and therefore its influence is relevant to single-year studies as well as those using panel data, we replicate our regression models year-by-year. Table 7 shows the mean coefficient values across years for each of the continuous variables, along with t-statistics formed using the time-series standard errors.

In Table 7, the model labeled I is identical to its counterpart in Table 5, except it is estimated one year at a time. The models labeled II and III in Table 7 correspond to models II

and III in Table 5, but without the year and industry indicator variables. Within years, the year indicator variable is of course collinear with the constant term. Our industry growth variable, *IndGrowth*, is identical for all firms in an industry within years, and so is collinear with the industry indicator variable. We replicated our results substituting industry indicators for *IndGrowth*, without changing any inferences.

Comparing the year-by-year results in Table 7 to the pooled results in Panel A of Table 5, we find a high degree of consistency in both the magnitudes of coefficient estimates and their statistical significance. An exception is the coefficient on volatility, which exhibits both lower magnitude estimates and reduced statistical significance in Table 7 as compared with Table 5. We interpret this as meaning that the volatility result in Table 5 is driven primarily by variation across years, not variation across firms within years. Given this, it is possible that our volatility measure, though measured firm-by-firm, primarily captures the effects of time-specific market volatility.

The high degree of consistency in the other results indicates that they are reliably measurable cross-sectional phenomena, and that the t-statistics in the panel regression have not been grossly exaggerated by pooling across years. In particular, we note that lagged analyst following has the same estimated coefficient values and retains strong statistical significance in the year-by-year setting. This indicates that it is not exclusively a time-series specification issue. The omitted variable problem exists for single-year models like Bhushan (1989) and Rock, Sedo and Willenborg (2000), and for year-by-year specifications like Alford and Berger (1999).

6. Summary and Concluding Remarks

In this study we have explored two specification issues in modeling analyst following. Using a sample of firms in five industries, we explore the effects of correcting for inertia in analyst coverage, and using ordered discrete response methods to estimate models of following. Using both within-sample statistical tests and out-of-sample predictions, we find that lagged analyst following is an important omitted variable in models of analyst following. We attribute

its importance to inertia in coverage decisions: the tendency of analysts to follow and neglect the same firms year to year. We find that controlling for the fact that analyst data are discrete by using ordered probit or logit produces a strongly significant benefit in terms of statistical fit, but a small benefit in terms of predictive accuracy, with questionable economic importance.

Our results suggest that, if the reason for modeling analyst coverage is to predict, then a simple first-order autoregression is probably as good as more complex models. Clearly, however, this simple model is not appropriate for testing competing theories about the determinants of analyst following. Our results suggest that for the latter purpose, including lagged analyst following along with other economic determinants is important for correctly specifying the model, and can change some inferences.

The contrast between strong statistical results and more modest predictive results in several of our comparisons illustrates the difference between statistical power and economic meaning. The large sample sizes available in this and related papers allow a high degree of statistical resolution. Economic significance, however, depends on the effects in the practical setting. Both the discreteness of analyst data and the predominance of low counts make a considerable difference in interpreting economic significance. Because analysts are indivisible, economic meaning attaches to whole individuals. Although we can in many cases achieve statistical significance equivalent to a few hundredths of an analyst, as a practical matter, only discrete increases of whole analysts matter. Our predictive results show that statistically measurable results can have quite modest economic significance.

We have studied the effects of inertia and discreteness in a sample of five industries: Food, Plastics, Computers, Lab Instruments, and Pharmaceutical/Biotech, over the ten-year period 1987-1996. The issues are pervasive in analyst data, both in other industries and in other time periods, so we believe our results will generalize to other samples. We selected relatively fast-growing industries, in the belief that this would maximize the range of analyst activity. To the extent that this criterion also selects industries that are less static in other causal factors, we conjecture that inertia will play an even larger role in other industries. We also selected

industries where earnings are relatively more value-relevant, in the belief that the standard measure of analyst following, existence of a current-year earnings forecast, should be more accurate in these industries. To the extent that this standard measure is less accurate in other industries, it would likely understate analyst following, with indeterminate implications for our results.

An avenue for future work is exploring the economic and behavioral underpinnings of analyst inertia. We propose that the auto-regressive nature of analyst following is more than a statistical artifact of an incomplete model, rather that it stems from structural characteristics of analysts' job. We conjecture that factors such as costs of learning and gathering information, and the ability to make company-specific links may come into play.[27] The setting in which to test these conjectures is in analyst-by-analyst data, not the firm-by-firm data we examine here.

[27] Regulation FD, recently enacted, may reduce the effect of analysts' company-specific links.

REFERENCES

ALFORD, ANDREW W., AND PHILIP G. BERGER (1999). A Simultaneous Equations Analysis of Forecast Accuracy, Analyst Following and Trading Volume, *Journal of Accounting, Auditing and Finance* 14 (Summer): 219-240.

BHUSHAN, RAVI (1989). Firm Characteristics and Analyst Following, *Journal of Accounting and Economics* 11: 255-274.

BOTOSAN, CHRISTINE A. (1997). Disclosure Level and the Cost of Equity Capital, *The Accounting Review* 72 (July): 323-350.

BARTH, MARY, RON KASZNIK AND MAUREEN MCNICHOLS (1999). Analyst Coverage and Intangible Assets, Working paper, Stanford University.

BRENNAN, MICHAEL J., AND PATRICIA J. HUGHES (1991). Stock Prices and the Supply of Information, *Journal of Finance* v. 46 n. 5: 1665-1691.

BRENNAN, MICHAEL J. AND AVANIDHAR SUBRAHMANYAM (1995). Investment Analysis and Price Formation in Securities Markets, *Journal of Financial Economics* 38: 361-381.

CAMPBELL, J. AND P. PERRON (1991). "Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots." *NBER Macroeconomics Annual*, 6, 141-201.

COCHRANE, D., AND G. ORCUTT (1949). Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms. *Journal of the American Statistical Association,* 44: 32-61.

DECHOW, PATRICIA M. (1994). Accounting Earnings and Cash Flows as Measures of Firm Performance: The Role of Accounting Accruals, *Journal of Accounting and Economics* v. 18 n. 1: 3-42.

DIEBOLD, F. AND M. NERLOVE, 1990.. Unit Roots in Economic Time Series: A Selective Survey. *Advances in Econometrics 8*, 3-69.

DUGAR, A. AND S. NATHAN, 1995. The Effects of Investment Banking Relationships on Financial Analysts' Earnings Forecasts and Investment Recommendations. *Contemporary Accounting Research* (Fall), 131-160.

FAMA, E. AND K. FRENCH (1988). "Permanent and Temporary Components of Stock Prices." *Journal of Political Economy*, 96, 246-273.

GREENE, W. *Econometric Analysis*, 1997. 3RD EDITION (NEW JERSEY: PRENTICE HALL).

HAYES, RACHEL M. (1998). The Impact of Trading Commission Incentives on Analysts' Stock Coverage Decisions and Earnings Forecasts, *The Journal of Accounting Research* 36 (Autumn): 299-320.

KENNEDY, PETER (1992). *A Guide to Econometrics* (3rd ed.). (Cambridge, MA: The MIT Press).

LANG, M. AND R. LUNDHOLM (1996). Corporate Disclosure Policy and Analyst Behavior, *The Accounting Review* 71 (October): 467-492.

LIN, H. AND M. MCNICHOLS (1998). Underwriting relationships, analysts' earnings forecasts and investment recommendations, *Journal of Accounting and Economics* 25 (February): 101-128.

LO, A. AND C. MACKINLAY (1988). "Stock market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test." *Review of Financial Studies*, 1, 41-66.

MCNICHOLS, MAUREEN, AND PATRICIA C. O'BRIEN (1997). Self-selection and Analyst Coverage, *Journal of Accounting Research* 35 (Supplement): 167-199.

MERTON, ROBERT C. (1987). A Simple Model of Capital Market Equilibrium with Incomplete Information, *Journal of Finance* 42 (July): 483-510.

MIKHAIL, MICHAEL B., BEVERLY R. WALTHER AND RICHARD H. WILLIS (1997). .Do Security Analysts Improve Their Performance with Experience?, *Journal of Accounting Research* 35 (Supplement): 131-157.

NELSON, C. R. AND C. I. PLOSSER (1982). "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications." *Journal of Monetary Economics*, 10, 139-162.

Nelson's Directory of Wall Street Research

O'BRIEN, PATRICIA C. AND RAVI BHUSHAN (1990). Analyst Following and Institutional Ownership, *Journal of Accounting Research* 28 (Supplement): 55-76.

POTERBA J. AND L. SUMMERS (1988). "Mean Reversion in Stock Returns: Evidence and Implications*, Journal of Financial Economics*, 22, 27-60.

RAJAN, RAGHURAM AND HENRI SERVAES (1997). Analyst Following of Initial Public Offerings, *Journal of Finance* v. 52, n. 2: 507-529.

ROCK, STEVE, STANLEY SEDO AND MICHAEL WILLENBORG (2000). Analyst Following and Count-data Econometrics, *Journal of Accounting and Economics* 30 (December): 351-374.

SCHWERT, G. W. (1987). "Effects of Model Misspecification on Tests for Unit Roots in Macroeconomic Data." *Journal of Monetary Economics*, 20, 73-103.

SCHMIDT, PETER (1976). *Econometrics*. (New York: Marcel Dekker, Inc.)

STOCK, J. AND M. WATSON (1986). "Does GNP Have a Unit Root?", *Economics Letters*, 22*,* 147-151.

VUONG, QUANG H. (1989). Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses, *Econometrica* v. 57, n. 2: 307-333.

Table 1
Sample selection, and distributon of observations across industries
and between estimation and hold-out samples

Panel A: Effects of data requirements on sample size, by industry[a]

Industry:	**Food**	**Plastics**	**Computers**	**Instruments**	**Biotech**	**All**
Obs. with trading volume during the fiscal year	1266	2137	515	1736	1247	6901
Obs. with complete data on all variables	1064	1868	456	1479	1118	5985
% lost to data requirements	16.0%	12.6%	11.5%	14.8%	10.3%	13.3%

Panel B: Estimation and holdout samples, by industry[b]

Industry:	**Food**	**Plastics**	**Computers**	**Instruments**	**Biotech**	**All**
Estimation sample	806	1416	339	1117	831	4509
Holdout sample	258	452	117	362	287	1476
Total	1064	1868	456	1479	1118	5985
Holdout sample, as a % of total	24.2%	24.2%	25.7%	24.5%	25.7%	24.7%

Table 1 (continued)
Sample selection, and distributon of observations across industries and between estimation and hold-out samples

Panel C: Number of observations, by level of analyst following and industry

Industry:	**Food**	**Plastics**	**Computers**	**Instruments**	**Biotech**	**All**	
Analyst Following[c]							**% of total**
0	406	721	251	508	528	2414	40.3%
1	100	234	64	183	145	726	12.1%
2	65	228	49	134	120	596	10.0%
3	57	149	30	95	81	412	6.9%
4	47	116	15	67	62	307	5.1%
5	40	66	7	59	24	196	3.3%
6-10	120	110	17	157	81	485	8.1%
11-15	67	76	11	106	45	305	5.1%
16-20	39	43	9	53	19	163	2.7%
21-25	50	18	1	43	11	123	2.1%
26-30	45	20	2	24	2	93	1.6%
31-35	28	24	0	22	0	74	1.2%
36-40	0	43	0	13	0	56	0.9%
41-45	0	17	0	12	0	29	0.5%
>46	0	3	0	3	0	6	0.1%
Total	**1064**	**1868**	**456**	**1479**	**1118**	**5985**	**100.0%**

Table 1 (continued)
Sample selection, and distributon of observations across industries and between estimation and hold-out samples

Notes:

[a]The data are observations drawn from five industries in the period 1987-1996. We use Compustat SIC codes to identify industries: 2010-2099 (Food), 3080-3089 (Plastics), 3570-3579 (Computers), 3820-3829 (Instruments), and 2830-2839 plus 8734 (Pharmaceutical and Biotech). We include a firm-year initially if it has trading volume on CRSP during the year. We narrow the sample to include only observations with complete data on all variables. We describe our variables in Table 3.

[b]We randomly assign 75% of observations to the estimation sample, and the remaining 25% to a hold-out sample. We generate random numbers using the SAS RANUNI procedure.

[c] Analyst following is the number of analysts listed on IBES during the fiscal year-end month.

Table 2
Observations by year and exchange, and by year and CRSP size decile[a]

Year	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
Panel A: number of observations, by exchange and year										
NYSE	89	89	93	96	99	104	106	103	112	96
AMEX	60	68	70	69	73	74	76	78	79	57
NASDAQ	392	368	370	360	375	414	463	494	538	520
	541	525	533	525	547	592	645	675	729	673
Panel B: percent of observations each year, by exchange										
NYSE	16.5	17.0	17.4	18.3	18.1	17.6	16.4	15.3	15.4	14.3
AMEX	11.1	13.0	13.1	13.1	13.3	12.5	11.8	11.6	10.8	8.5
NASDAQ	72.5	70.1	69.4	68.6	68.6	69.9	71.8	73.2	73.8	77.3
	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 2
Observations by year and exchange, and by year and CRSP size decile

Year	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
Panel C: percent of observations each year, by CRSP size decile										
decile 1	7.4	8.8	9.2	7.2	6.6	8.1	7.3	9.6	9.1	6.7
2	10.5	8.8	9.2	12.2	9.1	9.1	9.6	10.8	9.5	9.4
3	10.7	11.4	11.4	11.1	9.0	10.1	11.2	11.1	8.8	10.7
4	9.6	9.0	11.4	10.3	8.2	8.5	10.4	13.3	10.7	11.0
5	12.8	13.3	11.8	9.5	8.4	7.4	9.5	9.8	12.8	13.1
6	10.4	10.3	8.6	11.2	10.8	12.2	11.3	11.0	10.6	9.8
7	11.3	11.1	9.8	7.8	13.4	14.0	12.7	7.7	9.9	12.2
8	7.8	8.4	8.6	10.9	11.5	10.5	9.6	8.3	9.2	10.3
9	8.1	7.6	8.6	7.1	9.0	6.9	5.6	6.7	6.7	5.8
decile 10	11.5	11.4	11.3	12.8	14.1	13.2	12.9	11.7	12.9	11.1
	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Note:

[a]The data are observations with complete data from five industries in the period 1987-1996. See Table 1 for the industry definitions.

Table 3
Median values of firm characteristics in 1991, by level of analyst following[a]

Analyst following[b]	# of firms[c]	Size[d]	Volume[e]	Volatility[f]	Market Share[g]	Price[h]	Lagged following[i]
0	233	10.1	8.7	0.05	0.01%	$3.75	0
1	67	11.3	9.2	0.05	0.01%	$9.38	1
2	50	11.5	9.4	0.04	0.03%	$13.00	1
3	25	11.3	10.1	0.04	0.07%	$11.50	2
4	26	11.9	10.2	0.03	0.11%	$18.50	3.5
5	16	12.4	10.5	0.03	0.12%	$16.81	4
6-10	40	12.7	10.9	0.03	0.19%	$17.31	7
11-15	33	13.8	11.8	0.03	0.19%	$29.25	14
16-20	19	14.3	12.0	0.02	0.42%	$32.38	17
21-25	14	15.4	12.7	0.02	0.90%	$35.94	22.5
26-30	6	15.9	13.5	0.02	4.95%	$55.13	30
31-35	8	16.5	13.1	0.02	3.31%	$67.13	33.5
36-40	6	16.5	13.5	0.02	4.92%	$80.56	40
41-45	3	17.0	13.5	0.02	6.34%	$84.00	42
>46	1	17.4	13.8	0.01	10.18%	$88.25	45

Notes:

[a] The data are drawn from five industries, as described in Table 1. The reported medians are from the combined estimation and hold-out samples.
[b] Analyst following is the number of analysts listed on IBES during the fiscal year-end month.
[c] # of firms is the number of firms in our combined estimation and holdout samples.
[d] Size is the natural logarithm of market capitalization at fiscal year-end.
[e] Volume is the natural logarithm of average daily trading volume during the fiscal year.
[f] Variability is the standard deviation of returns over the fiscal year.
[g] Market share is company sales, as a % of industry sales.
[h] Price is stock price per share at fiscal year-end.
[i] Lagged following is the number of analysts listed on IBES during the fiscal year-end month of the prior year.

Table 4
Pairwise correlations of determinants of analyst following in 1991[a]

	Analyst following	Lagged following	Size	Volume	Volatility	Market Share	Price	Industry Growth
Analyst following	100.00%	96.44% 0.0001	78.31% 0.0001	65.74% 0.0001	-37.66% 0.0001	52.21% 0.0001	66.03% 0.0001	6.51% 0.1894
Lagged following	89.49% 0.0001	100.00%	75.22% 0.0001	61.23% 0.0001	-36.24% 0.0001	51.06% 0.0001	62.68% 0.0001	4.03% 0.4174
Size	73.14% 0.0001	66.37% 0.0001	100.00%	74.47% 0.0001	-45.95% 0.0001	46.88% 0.0001	75.61% 0.0001	23.61% 0.0001
Volume	65.81% 0.0001	55.28% 0.0001	73.63% 0.0001	100.00%	-15.94% 0.0012	33.12% 0.0001	43.15% 0.0001	26.91% 0.0001
Volatility	-48.42% 0.0001	-48.67% 0.0001	-48.15% 0.0001	-17.70% 0.0003	100.00%	-23.86% 0.0001	-44.18% 0.0001	-0.87% 0.8602
Market Share	55.68% 0.0001	61.51% 0.0001	45.06% 0.0001	23.49% 0.0001	-52.90% 0.0001	100.00%	45.07% 0.0001	-6.23% 0.2090
Price	67.40% 0.0001	57.64% 0.0001	81.97% 0.0001	52.00% 0.0001	-59.41% 0.0001	41.02% 0.0001	100.00%	20.77% 0.0001
Industry Growth	6.12% 0.2173	-1.13% 0.8198	31.08% 0.0001	31.29% 0.0001	-3.68% 0.4589	-40.41% 0.0001	27.66% 0.0001	100.00%

Notes:

[a] The data are drawn from five industries, as described in Table 1. See Table 3 for variable definitions. The correlations reported above are based on the 408 estimation sample observations in 1991. Pearson correlations and p-values appear above the primary diagonal. Spearman correlations and p-values appear below the primary diagonal. We truncate p-values at .0001.

Table 5
Regression and probit models of analyst following, 1987-1996[a]

Panel A: Estimated coefficients based on 4509 observations (with t- or z-statistics)[b]

	Pred.	Regression			Probit		
Variable[c]	**sign**	**I**	**II**	**III**	**I**	**II**	**III**
Size	+		1.15	0.25		0.39	0.24
			(17.7)	(8.7)		(20.8)	(11.9)
Volume	+		1.17	0.30		0.42	0.33
			(19.7)	(11.6)		(23.7)	(17.0)
Volatility	+		-1.11	-1.27		-17.88	-13.18
			-(0.4)	-(1.1)		-(13.9)	-(9.6)
Market Share	+		41.30	-0.25		-0.14	-2.36
			(12.4)	-(0.2)		-(0.2)	-(2.7)
Price	-		0.08	0.02		0.01	0.01
			(16.7)	(9.9)		(5.4)	(4.9)
Industry growth	+		0.45	0.66		-0.08	0.12
			(0.7)	(2.3)		-(0.5)	(0.6)
Lagged following	+	0.96		0.85	0.40		0.34
		(253.5)		(141.0)	(56.4)		(44.4)
test: $\alpha_9 = 1.0$		(10.8)		(25.7)	(85.8)		(85.8)

Table 5 (continued)
Regression and probit models of analyst following, 1987-1996

Panel B: F- and X^2-statistics on groups of catgorical variables based on 4509 observations[d]

	Regression			Probit		
	I	**II**	**III**	**I**	**II**	**III**
Year		39.8	12.2		206.7	72.6
p-value		0.0001	0.0001		0.0001	0.0001
Industry		5.7	2.8		20.4	7.5
p-value		0.0002	0.0267		0.0004	0.1126
Exchange		95.7	31.5		136.7	139.4
p-value		0.0001	0.0001		0.0001	0.0001

Panel C: Model summary statistics based on 4509 observations

	Regression			Probit		
	I	**II**	**III**	**I**	**II**	**III**
adj./pseudo R^2	93.44%	78.17%	94.42%	36.91%	28.60%	44.32%
F or X^2 statistic[e]	64253.5	494.9	3470.5	6746.8	5229.0	8101.5
p-value	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

Table 5 (continued)
Regression and probit models of analyst following, 1987-1996

Panel D: Holdout sample prediction errors based on 1476 observations[f]

	Regression			Probit		
	I	**II**	**III**	**I**	**II**	**III**
Criterion[g]						
MSE	4.33	18.82	3.71	4.37	10.69	3.41
RMSE	2.1	4.3	1.9	2.1	3.3	1.8
MA%E	38.8%	133.6%	40.0%	45.1%	60.6%	34.5%

Notes:

[a] The estimated models are as follows:

I: $$Analysts_{it} = a_0 + a_9 Analysts_{it-1} + \varepsilon_{1it}$$

II: $$Analysts_{it} = a_{0t} + a_{1industry} + a_{2exchange} + a_3 Size_{it} + a_4 Volume_{it} + a_5 Volatility_{it} + a_6 MktShare_{it} + a_7 Price_{it} + a_8 IndGrowth + \varepsilon_{1it}$$

III: $$Analysts_{it} = a_{0t} + a_{1industry} + a_{2exchange} + a_3 Size_{it} + a_4 Volume_{it} + a_5 Volatility_{it} + a_6 MktShare_{it} + a_7 Price_{it} + a_8 IndGrowth + a_9 Analysts_{it-1} + \varepsilon_{1it}$$

[b] We report the regression coefficient estimates with associated t-statistics in parentheses, and the probit coefficient estimates with associated z-statistics.
[c] See Table 3 for variable definitions.
[d] We report F-statistics on groups of categorical variables for the regression models, and X^2 statistics for the probit models. These statistics test the null hypothesis that the categories are indistinguishable.
[e] We report F-statistics for the regression models, and X^2 statistics for the probit models. These statistics test the null hypothesis that the model variables jointly have no explanatory power.
[f] The holdout sample is 25% of our original observations selected at random.
[g] The criteria are:
MSE = mean squared prediction error
RMSE = root mean squared prediction error
MA%E = mean of (absolute prediction errors, scaled by 1+lagged analyst following)

Table 6
Vuong test comparing regression and probit models of analyst following, 1987-1996, based on 4509 observations

	Model[a]		
	I	**II**	**III**
ln(likelihood) from Regression	-9,458.7	-12,854.9	-9,209.1
ln(likelihood) from Probit	-5,767.2	-6,526.1	-5,089.9
Vuong test statistic[b]	-47.8	-65.2	-46.2

Notes:

[a] The estimated models are as follows:

I: $$Analysts_{it} = a_0 + a_9 Analysts_{it-1} + \varepsilon_{1it}$$

II: $$Analysts_{it} = a_{0t} + a_{1industry} + a_{2exchange} + a_3 Size_{it} + a_4 Volume_{it} + a_5 Volatility_{it} + a_6 MktShare_{it} + a_7 Price_{it} + a_8 IndGrowth + \varepsilon_{1it}$$

III: $$Analysts_{it} = a_{0t} + a_{1industry} + a_{2exchange} + a_3 Size_{it} + a_4 Volume_{it} + a_5 Volatility_{it} + a_6 MktShare_{it} + a_7 Price_{it} + a_8 IndGrowth + a_9 Analysts_{it-1} + \varepsilon_{1it}$$

In this table, we use the categorical value of the dependent variable for both the regression and the probit estimations.

[b] The Vuong statistic tests the null hypothesis that the regression and probit estimations have the same likelihood. It has a limiting N(0,1) distribution. Negative values reject the regression model in favor of the probit model.

Table 7
Mean coefficients from year-by-year regressions of analyst following, 1987-1996[a]
(t-statistics in parentheses)

Variable[b]	Pred. sign	Regression I	II	III
Size	+		1.03	0.22
			(6.2)	(6.7)
Volume	+		1.29	0.32
			(10.7)	(10.2)
Volatility	+		-5.28	-3.11
			-(1.1)	-(2.4)
Market Share	+		35.74	-1.38
			(6.9)	-(0.6)
Price	-		0.11	0.03
			(5.3)	(5.5)
Industry growth	+		-0.74	-0.27
			-(1.5)	-(1.0)
Lagged following	+	0.96		0.85
		(60.1)		(35.4)

Notes:

[a] The estimated models are as follows:

I: $$Analysts_{it} = a_0 + a_9 Analysts_{it-1} + \varepsilon_{1it}$$

II: $$Analysts_{it} = a_0 + a_{2exchange} + a_3 Size_{it} + a_4 Volume_{it} + a_5 Volatility_{it} + a_6 MktShare_{it} + a_7 Price_{it} + a_8 IndGrowth + \varepsilon_{1it}$$

III: $$Analysts_{it} = a_0 + a_{2exchange} + a_3 Size_{it} + a_4 Volume_{it} + a_5 Volatility_{it}$$

[b] See Table 3 for variable definitions.