

# The kinks of financial journalism<sup>\*</sup>

Diego García<sup>†</sup>

Kenan-Flagler Business School  
University of North Carolina at Chapel Hill

October 20, 2014

## ABSTRACT

This paper studies the content of financial news as a function of past market returns. As a proxy for media content we use positive and negative word counts from general financial news columns from the Wall Street Journal and the New York Times. Our empirical analysis allows us to discriminate between theories that predict hyping good stock performance to those that emphasize negative news. The evidence is conclusive: negative market returns taint the ink of typewriters, while positive returns barely do.

*JEL classification:* G01, G14.

*Keywords:* media content, stock returns, journalism.

---

<sup>\*</sup>I would like to thank Paolo Fulghieri, David Hirshleifer, Camelia Kuhn, Geoff Tate and seminar participants at UNC at Chapel Hill, the University of Utah, the University of Luxembourg, Cheung Kong Graduate School of Business, the Shanghai Advanced Institute of Finance, the University of California at Irvine and the Tuck School of Business for comments on an early draft.

<sup>†</sup>Kenan-Flagler Business School, University of North Carolina at Chapel Hill, McColl Building, C.B. 3490, Chapel Hill, NC, 27599-3490. Tel: 1-919-962-8404; Fax: 1-919-962-2068; Email: [diego-garcia@unc.edu](mailto:diego-garcia@unc.edu); Webpage: <http://www.unc.edu/~garciadi>

# 1 Introduction

This paper studies the content of over seventy thousand columns on financial affairs from the New York Times and the Wall Street Journal from 1905 to 2005. In particular, we analyze journalists' word choices as a function of market events (DJIA returns). While focusing on a narrow functional area, we stand a good chance to understand what drove the word choice of journalists as a function of the impulses in their environment. Our large time-series sample allows us to see when journalists emphasize the positive domain (“hyping good news”) versus the negative domain (“darker descriptions” of stock market returns). Behavioral theories, as in Shiller (2000), argue that both are likely to be present, as the journalists look for ways to entertain the readers. Other theories suggest an informational role of journalists, in which both large positive and large negative market returns should play a role.<sup>1</sup>

Our empirical approach models the fraction of positive and negative words used by journalists as a function of market returns over the recent past. Clearly, lagged DJIA returns are excellent predictors of journalists' word choice. A parsimonious model with only lagged market returns can explain more than 30% of the variation in media content. More importantly, the effect is highly non-linear: positive returns have a much smaller impact on content than negative returns.

Figure 1 plots estimates of content from a non-linear model of the content of financial news, estimated with different linear functions over positive and negative lagged DJIA returns.<sup>2</sup> There is a very strong correlation between DJIA returns and content: higher market returns lead to journalists using more positive words, and less negative words. But there are a few kinks: (1) the slope over the negative domain is significantly steeper; (2) high positive DJIA returns barely influence media content, while high negative DJIA returns do.

The most robust feature of Figure 1 is the strong non-linearity over the domain of losses and gains. While the curves are rather steep around zero, the slope is higher on the domain of

---

<sup>1</sup>For example, theories of markets for information (e.g., Admati and Pfleiderer, 1986) would predict linear relationships between signals and content. The setting we have here concerns “public” news, namely lagged market returns. Throughout most of the sample such “public” signals were private to those with access to the market, since most of the investment public would receive this information via newspapers with a significant time lag.

<sup>2</sup>The fitted curves control for several other determinants of media content which we discuss at length in the modelling section.

losses than in the domain of gains. Furthermore, journalists barely make much of a difference between good (+1%) days and very good (+3%) days, whereas they do play up the difference between bad (-1%) days and really bad (-3%) days. This effect is even more pronounced for lagged returns two through four days before the writing of the article. We document that the asymmetry of journalists' writing is pervasive: it has barely changed from the 1920s to the 1990s, and virtually all authors exhibit the same pattern, emphasizing negative returns, ignoring large positive market moves.

The observational sample studied in this paper cannot rule out informational models that tie asset returns and words versus behavioral arguments for the existence of the non-linearities. Nonetheless, our evidence does challenge some theories. The linearity of many informational economics models, based on CARA preferences and Gaussian payoffs, is called into serious question. Our results suggest that the “hyping” in Shiller (2000) cannot generally be associated with higher lagged market performance.

The evidence is consistent with models of human behavior where the domain of gains and losses trigger different reactions by agents, the journalists, whereas it is inconsistent with models where information flows in a continuous manner from fundamentals to the printed page. The concavity in the domain of gains and the convexity in the domain of losses fits well with Kahneman and Tversky (1979)'s prospect theory. The jump at zero is very suggestive of the existence of a reference point effect at the point of gains/losses.

The data under scrutiny is a particularly nice laboratory. Financial news as a discipline is a well documented profession that goes back well into the 19th century. Further, the sources of discussion (earnings, stock price movements) are also easily measurable, and by and large, they are *continuous* variables, i.e. DJIA returns. The columns we study are coloured around such variables, which gives us a particularly sharp test of the impulse response function from market returns to the printed page.<sup>3</sup> The sources we include in our analysis were the leading media providers of financial news in the United States for the majority of our sample period, and, as such, provide a particularly nice setting in which to study how journalists report financial news.

---

<sup>3</sup>One can consider our empirical study a nice laboratory where to look at impulse-response “in the field” more generally.

Most of the literature studying the media in economics has focused on the effect of media on economic outcomes.<sup>4</sup> This contrasts with the questions in this paper, which studies the impact of economic variables (stock returns) on the content of the news. Many papers study predictors of media content mostly in order to create instruments to argue causality in a second stage.<sup>5</sup> In the journalism literature, there are several studies of content as a function of economic variables. Bow (1980) argues that there were no predictive signs in the media prior to the 1929 stock market crash, while Griggs (1963) gives a similar account in the context of the 1957–1958 recession. Neilson (1973) discusses the state of journalism during the bulk of our sample. Norris and Bockelmann (2000) and Roush (2006) also have extensive discussions as to the role of the media since the beginning of the 20th century.

The rest of the paper is structured as follows. In Section 2 we introduce the data we use in our empirical analysis. Section 3 discusses our main results, while Section 4 looks at author fixed effects, different time periods, lower-frequency returns, as well as other indexes. Section 5 concludes.

## 2 The data

The paper uses several sources of data. The first is stock return information. For the majority of our analysis, we use the Dow Jones Industrial Average from Williamson (2008).<sup>6</sup> The Dow Jones Industrial index goes back to the turn of the 20th century, and thus allows us to have a metric of US stock returns prior to the coverage in the more standard Center for Research in Security Prices (CRSP), which started in 1926. We let  $R_t$  denote the log-return on the DJIA index on date  $t$ . Business cycle information is obtained from the NBER website

---

<sup>4</sup>Media slant has been a topic of interest in politics (Gentzkow and Shapiro, 2006, 2008), as well as in finance (Cutler, Poterba, and Summers, 1989; Klibanoff, Lamont, and Wizman, 1998; Chan, 2003; Barber and Odean, 2008; Gaa, 2008; Tetlock, Saar-Tsechansky, and Macskassy, 2008; Yuan, 2008; Engelberg, 2008; Fang and Peress, 2009; Solomon, 2012; Bhattacharya, Galpin, Yu, and Ray, 2009).

<sup>5</sup>For example, Engelberg (2008) and Gurun and Butler (2012) use geography to predict content, Dougal, Engelberg, García, and Parsons (2012) use author’s identity. See Reuter and Zitzewitz (2006) for evidence on bias created by economic ties.

<sup>6</sup>Historical data is available free of charge from <http://www.djaverages.com/>, including the total return for the Dow Jones Industrial Average, but this source does not include Saturday data. For this reason, we use the data on the DJIA from Williamson (2008), see <http://www.measuringworth.org/DJA/>. Exclusion of the Saturday data does not affect any of our results.

<http://www.nber.org/cycles.html>. The last source of data is the media content of three different columns of financial news, which we describe next.

The media content measures are constructed starting from the Historical New York Times Archive and the Historical Wall Street Journal Archive. The former goes back to the origins of the newspaper in 1851, while the latter starts in 1889. These datasets were built by scanning the full content of the newspaper, cropping columns separately. In order to have a consistent set of articles that cover financial news, we focus on three columns that were published daily during this period. From the New York Times we use the “Financial Markets” column, and the “Topics in Wall-Street” column (García, 2013), and from the Wall Street Journal we use the “Abreast of the market” column (Tetlock, 2007). The “Topics in Wall-Street” column ran daily under different titles (i.e. “Sidelights from Wall-Street”, “Financial and Business Sidelights of the Day,” “Market Place”) until the end of our sample period. The “Financial Markets” column stopped being published with such a heading in the 1950s, although the New York Times obviously continued to publish a column with the financial news for the day, which we use in our analysis. The “Abreast of the Market” column was published daily virtually uninterrupted from 1926–2007, see Tetlock (2007) for details. The paper studies 76,537 pdf files from the Historical Archives that were associated with either of these columns from January 1, 1905 through December 31, 2005.<sup>7</sup> A total of 55,168 of the columns in our sample were from the New York Times, while the other 21,369 are from the Wall Street Journal.

The columns under study were essentially summaries of the events in Wall Street during the previous trading day. The average article had around 800 words. The articles discussed anything from particular companies or industries to commodities and general market conditions. The topics included in the columns were of a business nature, with a focus on financial matters. Tetlock (2007) and García (2013) give more detailed accounts of the data sources.

To construct the media content measures, we transform the scanned images available from the New York Times Historical Archive into text documents. This is referred to in the computer science literature as “optical character recognition” (OCR). We use ABBYY software, the leading

---

<sup>7</sup>We exclude news from the period in 1914 when the NYSE was closed (up to December 12th, 1912).

package in OCR processing, to convert the images into text files. Although the quality of the transcription of the articles is high, it is important to notice that the accuracy of OCR processing may be low for some files. The quality of the scanned images in the NYT Historical Archive is particularly low prior to 1905, thus our choice of starting date.<sup>8</sup> We note that this approach to reading text only adds random noise to our media content measures, and thus it will not bias our conclusions.

In order to quantify the content of the New York Times articles, this paper takes a “dictionary approach.”<sup>9</sup> For each column  $i$  written on date  $t$ , we count the number of positive words,  $g_{it}$ , and negative words,  $b_{it}$ , using the word dictionaries provided by Bill McDonald.<sup>10</sup> As argued in Loughran and McDonald (2011), standard dictionaries fail to account for the nuances of Finance jargon, thus the categorization we use has particular merits for processing articles on financial events. We let  $w_{it}$  denote the total number of words in an article. We construct these media measures dating them to the day  $t$  in which they were written, with the understanding that they are published in the morning of day  $t + 1$ . The rationale is that the information contained in these columns clearly belongs to date  $t$ . The writing process for each article started at 2:30-3:00 pm, typically just as the market was about to close, and the final copy was turned in to be edited and typeset at around 5-6 pm.

We aggregate the media content measures to create a time-series that matches the Dow Jones index return data available. In particular, we first combine all news printed between two trading dates, in order to be conservative with our standard errors, and also reduce noise stemming from particular idiosyncrasies from each column. In essence, we are trying to measure the content of the financial news on investors’ desks prior to the opening of the market, and model its relationship to previous market events. We will also study cross-sectional variation with respect to each column, since the Sunday and Monday columns are likely rather different, both in terms

---

<sup>8</sup>The OCR software will try to interpret anything in the original image, from spots to actual text. Different margins, multiple columns, and page formatting issues in general present a challenge for the character recognition process.

<sup>9</sup>Non-dictionary approaches have gained much popularity in recent research on text content analysis, in which not just the words, but the order and their role in a sentence is taken into account (i.e. the Diction software used in Demers and Vega, 2008). Given the OCR processing issues discussed above, these types of language processing algorithms are not appropriate for our study.

<sup>10</sup>See [http://www.nd.edu/~mcdonald/Word\\_Lists.html](http://www.nd.edu/~mcdonald/Word_Lists.html) for details.

of circulation and material on which to write about, from the other weekday columns.

In order to aggregate the news, we average the measures of positive/negative content from articles that were written since the market closed until the market next opens. When the market is open on consecutive days,  $t$  and  $t + 1$ , we define our daily measure of positive media content as  $G_t = \sum_i g_{it} / \sum_i w_{it}$ , where the summation is over all articles written in date  $t$  (given our news selection, there are two such articles for the majority of days in our sample). Similarly, we construct our daily measure of negative media content as  $B_t = \sum_i b_{it} / \sum_i w_{it}$ . In essence, we count the number of positive and negative words in the financial news under consideration, and normalize them by the total number of words. For non-consecutive market days we follow a similar approach, including all articles published from close to open. To be precise, consider two trading days  $t$  and  $t + h + 1$  such that  $h > 0$  and the market was closed  $h$  days, from  $t + 1$  through  $t + h$ . We define the positive media content measure as  $G_t = \sum_{i,s=t}^{s=t+h} g_{is} / \sum_{i,s=t}^{s=t+h} w_{sh}$ . We proceed analogously for the negative media content variable and define  $B_t = \sum_{i,s=t}^{s=t+h} b_{is} / \sum_{i,s=t}^{s=t+h} w_{sh}$ . We define the pessimism factor as the difference between the negative and positive media content measures, i.e.  $P_t = B_t - G_t$ .

For consecutive trading dates, our media measures  $G_t$  and  $B_t$  are constructed using information that was available as of the end of date  $t$  when the market is open on date  $t + 1$  (the bulk of our sample). It is less clear whether market prices on date  $t$  reflected the information available to the journalists writing the columns, as the deadline for turning in the article to the editor was not until roughly 5-6pm, while the NYSE closed at 3-4pm. We further remark that for non-consecutive trading dates, we use articles that may have been written on days after date  $t$ , but prior to the market opening (i.e., in the case of holidays).

Table 1 presents summary statistics on our media measures. The average number of positive and negative words, averaged over all articles, are 1.27% and 2.08% respectively. The articles from the Wall Street Journal use slightly higher fraction of positive words, 1.42 versus 1.21, but the fraction of negative words are virtually identical, 2.08 versus 2.09. We remark how our time-series aggregate, which adds all articles between trading dates, have similar means, as expected, since they just weight the different articles by their word length. The standard

deviation of this time series aggregate is significantly less noisy. For example, positive content has a standard deviation from 0.36 when aggregated, versus 0.63 when looking at individual articles. The media content variable, which simple subtracts negative from positive frequencies, inherits the properties just discussed for their individual components.

For the rest of the paper we normalize our sentiment measures so they have zero mean and unit variance. This will allow us to interpret the regression coefficients in terms of one-standard deviation shocks to the sentiment measures, thus making it easier to gauge the economic magnitude of our results.

### 3 Media content and DJIA returns

We start by estimating a parsimonious time-series model of media content. In particular, we assume the following econometric specification:

$$M_t = \beta_0 R_t + \beta \mathcal{L}(R_t) + \rho \mathcal{L}(M_t) + \eta X_t + \epsilon_t; \quad (1)$$

where  $M_t$  denotes the media content written between trading dates  $t$  and  $t + 1$  (for articles written after the market closed on date  $t$  by prior to opening on date  $t + 1$ ), and  $R_t$  denotes the log-return on the DJIA on date  $t$  (from close on date  $t - 1$  to close on date  $t$ ). We truncate  $R_t$  at  $-3\%$  from below and  $+3\%$  from above in the analysis that follows. The set of explanatory variables  $X_t$  includes day-of-the-week dummies and a cubic function of time.

Table 2 presents estimates of (1). The first set of columns (“Only  $\mathcal{L}(R_t)$ ”) present estimates under the constraint  $\rho = 0$ , while the second set of columns (“Only  $\mathcal{L}(M_t)$ ”) present estimates under the constraint  $\beta = 0$ . The last two columns present the unconstrained estimates. The  $t$ -statistics reported in the table use Newey-West corrections with ten lags.

The most important determinant of media content is the last trading day returns, as expected. A one-standard deviation shock to returns moves media content by one-half of a standard deviation, a large effect in economic terms. The statistical significance is also very large, due to the large sample we are studying, and the strong correlation between  $M_t$  and  $R_t$ . Perhaps



more surprising is the fact that lagged returns, even nine days into the past, have significant predictive power. The economic magnitudes decline quickly with the distance between the return date and the writing date, but the aggregate effect of returns lagged 5–9 days is non-trivial. Overall, lagged returns explain 34.6% of the variation in media content.

The second column presents the estimates ignoring lagged returns, but including lagged media content. The autocorrelation structure given in columns 4-5 of Table 2 shows that  $M_t$  is a fairly persistent process. But lags of media content can only explain 22.4% of the variation in media content itself. The last two columns give the unconstrained model. We highlight how the introduction of lagged media content does increase the  $R^2$  of the regression, from 34.6% to 42.3%. The autocorrelation of media content is 0.141 in this specification, which suggests there is a persistent component, but it is not very large. This persistence can be easily explained by author fixed effects, for example, as the same journalists would write the columns at hand during different periods of time, each with their own style (Dougal, Engelberg, García, and Parsons, 2012).

It is important to note that the coefficient on  $R_t$  is still virtually unchanged, 0.478 versus 0.470. Furthermore, the impulse response of  $M_t$  to  $R_{t-k}$  is also not different from that implied by the first set of columns to the last set. For example, the impulse response to  $R_{t-1}$  is  $\beta_1 = 0.149$  in the first specification, and  $\beta_1 + \rho_1\beta_0 = 0.143$  in the last. The impulse response to  $R_{t-2}$  is  $\beta_2 = 0.082$  in the first specification, and  $\beta_2 + \rho_1\beta_1 + \rho_1^2\beta_0 + \rho_2\beta_0 = 0.076$  in the last.

Our main specification to capture non-linearities will consist of a model of media content of the form:

$$M_t = f(R_t; \alpha, \beta) + \eta X_t + \epsilon_t; \tag{2}$$

where  $M_t$  denotes the media content written between trading dates  $t$  and  $t + 1$  (for articles written after the market closed on date  $t$  by prior to opening on date  $t + 1$ ), and  $R_t$  denotes the log-return on the DJIA (from close of date  $t - 1$  to close of date  $t$ ), truncated at  $-3\%$  from below and  $+3\%$  from above. The set of explanatory variables  $X_t$  includes 10 lags of  $R_t$  and  $M_t$ , as well as day-of-the-week dummies and a cubic function of time.

Since our main interest is on potential non-linearities between the outcome variable, media content, and economic variables, we propose a parsimonious, yet flexible function  $f(R_t; \alpha, \beta)$ . In particular, we assume that  $f$  is of the form

$$f(R_t; \alpha, \beta) = \sum_{i=1}^4 (\alpha_i + \beta_i R_t) 1_{R_t \in S_i} \quad (3)$$

where the sets  $S_i$  are:  $S_1 = (-3, -1)$ ,  $S_2 = (-1, 0)$ ,  $S_3 = (0, 1)$ , and  $S_4 = (1, 3)$ .

We also estimate a model with “smooth” non-linearities, which we use in our plots. While such a model cannot estimate the jump at zero, it does corroborate our main parametric conclusions.<sup>11</sup> The choice of the set of intervals  $S_i$  is actually motivated by the fit of a model based on splines. The fact that we impose a linearity restriction for each interval makes hypothesis testing and the interpretation of coefficients simply more transparent.

Table 3 presents the estimates of the model in (2). We focus first on the slope coefficients,  $\beta_i$ , given in the right. Not surprisingly, the content of financial news is in large part determined by the market returns during the last trading day (proxied by the DJIA). All parameters  $\beta_i$ ,  $i = 1, \dots, 4$ , are significant at standard levels of significance. Two differences stand out: (1) the slope in the negative “normal” range  $(-1, 0)$  is  $\beta_2 = 0.656$ , whereas that in the “normal” positive range  $(0, 1)$  is only  $\beta_1 = 0.528$ . While both are significantly higher than the slopes at “tail” market returns (above 1% or below -1%), the slope on the negative domain is significantly higher than on the positive domain. The test on Panel B yields a Newey-West adjusted test with a  $p$ -value well below 1%.<sup>12</sup>

When the DJIA rose by more than 1%, the news content becomes significantly less sensitive to market returns. The point estimate of  $\beta_4 = 0.057$  is statistically significant, but small in economic terms. Moving DJIA returns from +1% to +3% changes the content of financial news by little more than 1/10th of a standard deviation. In contrast, in the negative “tail” domain, the coefficient  $\beta_1 = 0.203$  tells us a similar market move would change media content by 4/10ths

---

<sup>11</sup>Estimating the model with splines that allow for a discontinuity at zero only reinforces our findings with respect to a jump at zero.

<sup>12</sup>All statistics reported use Newey-West corrections with ten lags.

of a standard deviation. As the test in the last row of Panel B documents, the difference is significant.

The tests of the slopes just discussed, both around zero and for larger market moves, shed some light on the differences in writing on the domain of gains and losses. We have rejected the null hypothesis of a function that is smooth in first derivatives at the reference point of 0% market returns. Our last test in this section studies whether there is a jump in the news content itself, that is, whether the function  $f$  is discontinuous at zero.

Given our parametric specification, we can test for such a jump simply by comparing the intercept coefficients  $\alpha_2$  and  $\alpha_3$ . Their difference is 0.152, which is non-trivial in economic terms, and highly statistically significant, as reported in Panel B. We conclude that there is indeed a difference between reporting very small positive returns, versus reporting very small negative returns. The boundary of the domain of gains and losses acts as a reference point for journalists.

In our next set of tests, we augment the specification in (2) by adding non-linear functions of DJIA returns two to four days before the publishing of the papers (measured on trading days time). This is a nice test of our previous results, since both Tetlock (2007) and García (2013) document that media content is influenced by lags going back at least four trading periods.<sup>13</sup> While there is little reason to suspect that the returns 2–4 days ago would have much of an effect as a “reference point,” the point estimates on the slope coefficients act as an alternative test of the “kinks” documented in Table 3 and Figure 1.

In particular, we estimate the model

$$M_t = f_1(R_t; \alpha_1, \beta_1) + f_2(R_{t-1}; \alpha_2, \beta_2) + f_3(R_{t-2}; \alpha_3, \beta_3) + f_4(R_{t-3}; \alpha_4, \beta_4) + \eta X_t + \epsilon_t;$$

where  $M_t$  denotes the media content written between trading dates  $t$  and  $t+1$  (for articles written after the market closed on date  $t$  by prior to opening on date  $t+1$ ),  $R_t$  denotes the log-return on the DJIA (truncated at  $-3\%$  from below and  $+3\%$  from above). The set of explanatory

---

<sup>13</sup>Other low-frequency variables, such as GDP, also influence the writing, but at the daily frequency we are working with, their influence is both economically and statistically small (the increase in explanatory power from those variables, compared to recent lagged market returns, is negligible).

variables  $X_t$  includes day-of-the-week dummies and a cubic function of time.<sup>14</sup> The functions  $f_j(R_t; \alpha, \beta)$  are assumed to be of the form

$$f_j(R_{t-j}; \alpha_j, \beta_j) = \sum_{i=1}^4 (\alpha_{ji} + \beta_{ji} R_{t-j}) 1_{R_{t-j} \in S_i} \quad (4)$$

where the sets  $S_i$  are:  $S_1 = (-3, -1)$ ,  $S_2 = (-1, 0)$ ,  $S_3 = (0, 1)$ , and  $S_4 = (1, 3)$ . In essence, we reproduce the results from the main specification allowing for non-linearities for all lagged returns  $R_t$  through  $R_{t-3}$ . As before, all statistics reported in the table use Newey-West corrections with ten lags.

Panel A.1 in Table 4 mimics the first panel in Table 3, the relationship between media content and the previous day DJIA returns, controlling for non-linearities with respect to lagged returns  $R_{t-2}$  through  $R_{t-4}$ . The estimates are virtually identical. In particular, the difference between the intercepts around zero is actually slightly larger (0.156 versus 0.154), and the differences between the slopes in the positive and negative domain are also larger (0.656 versus 0.528 before, 0.699 and 0.516 now, around zero; 0.240 and 0.027 now, 0.203 and 0.057 before). Thus the previous results are not affected by a slightly difference econometric specification.

Panel A.2 in table 4 studies potential non-linearities with respect to the returns two trading days before publication. Looking at the “Intercepts” column, we see that there is no difference in the intercepts  $\alpha_{22}$  and  $\alpha_{23}$ , i.e. there is no “jump at zero” with respect to the return two days ago. In fact, all the intercepts around zero are very close to each other ( $-0.051$  and  $-0.053$ ,  $0.009$  and  $-0.006$ ,  $-0.003$  and  $0.040$ ), in contrast with the results for the returns the previous trading day.

Looking at the slope coefficients in Panel A.2, we see that media content increases with stock returns two trading days ago, but most notably in the negative domain (the sets  $S_1$  and  $S_2$ ). Remarkably, this result, which showed strongly in Table 3, comes out as strong in the regressions with the DJIA returns lagged two through four days. This is apparent in Panel A.2,

---

<sup>14</sup>We do not include  $\mathcal{L}(M_t)$  in (3) in order to avoid computing impulse response functions using our non-linear specification, a non-trivial task. The results in Table 2, discussed at the beginning of this section, suggest such omission should not bias our conclusions.

where the point estimates for the slopes around zero are 0.248 (negative domain) and 0.109 (positive domain), compared to 0.699 and 0.516 for the one-day lag. While they are smaller, as expected, their difference is large in economic terms. The same pattern emerges for 3-days and 4-days lagged returns, the slopes around zero are 0.155 and 0.099 (negative domain) and 0.047 and 0.006 (positive domain).

While the statistical power at the “tails” is smaller, Panel A.2 of Table 4 documents very strong differences in the intervals  $(-3, -1)$  and  $(1, 3)$  for the returns from two days ago. The slope coefficient in the negative domain is 0.185, versus only 0.025 in the positive domain. The slopes at the “tails” for the returns three and four days before writing exhibit a similar pattern: the slope in the negative domain is positive and statistically different from zero at standard levels of confidence, whereas the slope in the positive domain is actually negative.

Panel B of Table 4 reports formal  $F$ -tests of one-dimensional restrictions, as in Table 3. All claims in the previous discussion hold. With respect to slope tests, all of them are highly significant with the exception of the tests with returns four days ago (highest  $p$ -value 2%). With respect to the intercept tests the only highly significant is the one that includes the last trading day, which is the only natural reference point.

## 4 Cross-sectional variation

In the previous section we established that the response of media content to lagged market returns was more pronounced in the negative domain than in the positive domain. In this section, we explore what may drive the kinks reported in section 3. We first look at variation of the authorship of columns, and also analyze to what extent the evidence provided in Table 3 is stable throughout our sample period. We also extend our analysis by looking at other indexes: instead of the DJIA, which comprises at most thirty companies, we use a value-weighted index, as well as indexes of small and large firms.<sup>15</sup> The last set of tests ask whether the asymmetries may have a time-series component that varies with the business cycle, and we estimate our main specification in recessions and expansions. We also study to what extent the parametric

---

<sup>15</sup>The indexes are from Ken French’s data library.

assumptions, in particular the functional form (3), affect our inferences.

### [DISCUSSION OF AUTHOR-FIXED EFFECTS, Table 5]

Table 6 replicates the analysis in Table 3, for four different time periods. If our findings were driven by supply side considerations (i.e. journalists peculiarities), we should find different estimates through our sample period, as both the editor of the financial section, and the team of journalists writing the stories, changed multiple times.<sup>16</sup>

The evidence in Table 6 suggests that the non-linearities previously reported appear throughout our sample period. The estimates for  $\beta_1$ , for example, which measures the slope in the left-tail of market returns, are 0.241, 0.255, 0.286, and 0.236, whereas those in the right-tail, measured by  $\beta_4$ , are 0.024, 0.067, 0.094, and 0.116, for the time periods 1905–1930, 1931–1955, 1956–1980, and 1981–2005 respectively. It is rather remarkable how stable the non-linearities from Table 3 are: in all four time periods we have that  $\beta_1 > \beta_4$ , and that  $\beta_2 > \beta_3$ , i.e. the reaction of news to market returns are significantly higher in the negative domain.

### [DISCUSSION OF LOW-FREQUENCY RETURNS, Table 7, Figure 3]

Table 8 replicates the analysis in Table 3, using three different indexes. Our previous results are virtually unchanged, if anything slightly larger in magnitude. The models all detect a jump at zero. Furthermore, the tail slopes on the positive domain are all below 0.06, and statistically insignificant. This is in contrast to the point estimates on the negative domain, which are all large, from 0.14 to 0.28, and statistically significant. There is evidence that large stocks matter more in terms of media content, as expected, but small stocks returns also help predict media content, with similar kinks to those reported in Table 3.

Our next set of tests asks whether the impulse responses we have documented vary along

---

<sup>16</sup>See Dougal, Engelberg, García, and Parsons (2012) for a study of the role of journalist fixed effects in financial news. Journalists started signing their articles in the published versions of the Wall Street Journal and the New York Times starting around 1970, and thus we cannot use author identity in our study. For the sample period 1970–2005 we find, in unreported results, similar results to those reported in Table 6: all authors exhibit the type of kinks plotted in Figure 1.

the business cycle. There are reasons to suspect that this may be the case, from marginal utility arguments to theories based on psychology and mood during good and bad times. Table 9 estimates two different non-linear functions, with the parametric representation in (3), one during expansions, one during recessions, using NBER definitions. Columns 2–3 in Table 9 present the results on expansions, whereas columns 4–5 include those for recessions. The non-linear fit from our main specification in Table 3 almost gets copied across the two sets of columns: There is little reaction to large positive returns, in contrast to large negative returns. Even the magnitudes are very similar in economic terms. We conclude that the kinks that this paper documents are a general pattern that does not hinge on particular market states.<sup>17</sup>

There were three tests conducted in Table 3: two tests regarding behavior in the “middle” ( $\alpha_2 = \alpha_3$  and  $\beta_2 = \beta_3$ ), and one on slopes at the tails ( $\beta_1 = \beta_4$ ). We mimic the analysis next by extending the number of linear splines. In particular, assume that the function  $f(R_t; \alpha, \beta)$  is of the form

$$f(R_t; \alpha, \beta) = \sum_{i=1}^8 (\alpha_i + \beta_i R_t) 1_{R_t \in S_i} \quad (5)$$

where the sets  $S_i$  are:  $S_1 = (-3, -2)$ ,  $S_2 = (-2, -1)$ ,  $S_3 = (-1, -0.5)$ ,  $S_4 = (-0.5, 0)$ ,  $S_5 = (0, 0.5)$ ,  $S_6 = (0.5, 1)$ ,  $S_7 = (1, 2)$  and  $S_8 = (2, 3)$ .

We start discussing the tests on the tails of market returns. Looking at the columns labelled “Slopes,” we find even stronger evidence of asymmetries in Table 10, compared to Table 3. Both  $\beta_1$  and  $\beta_2$  are positive, and significantly different than zero. On the other hand,  $\beta_7$  is barely positive, and  $\beta_8$  is actually negative. The fact that there is no “hyping” of large positive returns is very robust, as is the fact that large negative returns do influence journalists’ word choices.

Turning to the behavior around zero market returns, we find that, in contrast with the results in Table 3, the slope around zero is slightly higher in the domain of gains than that of losses,  $\beta_5 > \beta_4$ , albeit the difference is not statistically significant ( $p$ -value 0.271). There is also no slope difference in the  $(-1, -0.5)$  and the  $(0.5, 1)$  ranges. In contrast, the model still detects a significant jump at zero.

---

<sup>17</sup>The NBER business cycle dummies are a good proxy for most other “market downturn” proxies one can empirically develop. In unreported results we find the non-linearities we document in Table 3 in time-series subsets sorted by lagged market returns and market volatility.

These findings hold across a large set of non-linear models. The slope on the positive domain just above the zero market returns point is rather steep (see also Figure 1). Inferences just around zero on slopes are mixed. But starting at  $-0.5$  and  $+0.5$  these differences start to be noticeable, and they become more pronounced as we move out in the gains and losses domains. The differences between the behavior in the right tail and the left tail, for any reasonable definition of tails, are both statistically and economically significant. At the same time, the jump at zero is surprisingly robust: only when tests are very narrow, on the range  $(-0.10, 0.10)$ , does the  $F$ -test become insignificant.

## 5 Conclusion

This paper has established a strong non-linearity between lagged market returns and the content of financial news. The shape of the relationship is present in all subsamples we have studied, and holds not just for the last trading day, but even for returns 2–4 days before publication. While the shape in Figure 1 resembles Kahneman and Tversky (1979)’s prospect theory, only a theoretical model of media production will provide some further insights into what drives our empirical findings. Our paper suggests such model needs to generate the kinks we observe in the writing of financial journalists.



## References

- Admati, A. R., and P. Pfleiderer, 1986, “A monopolistic market for information,” *Journal of Economic Theory*, pp. 400 – 438.
- Barber, B., and T. Odean, 2008, “All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors,” *Review of Financial Studies*, 21(2), 785–818.
- Bhattacharya, U., N. Galpin, X. Yu, and R. Ray, 2009, “The role of the media in the internet IPO bubble,” *Journal of Financial and Quantitative Analysis*, 44, 657–682.
- Bow, J., 1980, “The “Times’s” Financial Markets column in the period around the 1929 crash,” *Journalism Quarterly*, 57, 447–450.
- Chan, W. S., 2003, “Stock price reaction to news and no-news: drift and reversal after headlines,” *Journal of Financial Economics*, 70, 223–260.
- Cutler, D. M., J. M. Poterba, and L. H. Summers, 1989, “What Moves Stock Prices?,” *Journal of Portfolio Management*, 15, 4–12.
- Demers, E. A., and C. Vega, 2008, “Soft Information in Earnings Announcements: News or Noise?,” working paper, INSEAD.
- Dougal, C., J. Engelberg, D. García, and C. Parsons, 2012, “Journalists and the stock market,” *Review of Financial Studies*, 25(4), 639–679.
- Engelberg, J., 2008, “Costly information processing: evidence from earnings announcements,” working paper, University of North Carolina.
- Fang, L. H., and J. Peress, 2009, “Media coverage and the cross-section of stock returns,” *Journal of Finance*, 64(5), 2023–2052.
- Gaa, C., 2008, “Good news is no news: asymmetric inattention and the neglected firm effect,” working paper, University of British Columbia.
- García, D., 2013, “Sentiment during recessions,” *Journal of Finance*, 68(3), 1267–1300.
- Gentzkow, M., and J. M. Shapiro, 2006, “Media bias and reputation,” *Journal of Political Economy*, 114, 380–316.
- , 2008, “Competition and truth in the market for news,” *Journal of Economic Perspectives*, 22, 133–150.
- Griggs, H., 1963, “Newspaper performance in recession coverage,” *Journalism Quarterly*, 40, 559–564.
- Gurun, U. G., and A. W. Butler, 2012, “Don’t believe the hype: local media slant, local advertising, and firm value,” *Journal of Finance*, 67, 561–597.

- Kahneman, D., and A. Tversky, 1979, "Prospect theory: an analysis of decision under risk," *Econometrica*, 47, 263–292.
- Klibanoff, P., O. Lamont, and T. Wizman, 1998, "Investor reaction to salient news in closed-end country funds," *Journal of Finance*, 53(2), 673–699.
- Loughran, T., and B. McDonald, 2011, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *Journal of Finance*, 66, 35–65.
- Neilson, W., 1973, *What's News – Dow Jones*. Clinton Book Company, Radnor, PA.
- Norris, F., and C. Bockelmann, 2000, *The New York Times — Century of Business*. McGraw-Hill, New York City, New York.
- Reuter, J., and E. Zitzewitz, 2006, "Do Ads Influence Editors? Advertising and Bias in the Financial Media," *Quarterly Journal of Economics*, 121(1), 197–227.
- Roush, C., 2006, *Profits and Losses*. Marion Street Press, Oak Park, Illinois.
- Shiller, R. J., 2000, *Irrational Exuberance*. Princeton University Press, Princeton.
- Solomon, D., 2012, "Selective publicity and stock prices," *Journal of Finance*, 67(2), 599–637.
- Tetlock, P. C., 2007, "Giving content to investor sentiment: the role of media in the stock market," *Journal of Finance*, 62(3), 1139–1168.
- Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy, 2008, "More than words: quantifying language to measure firms' fundamentals," *Journal of Finance*, 63(3), 1437–1467.
- Williamson, S. H., 2008, "Daily Closing Values of the DJIA in the United States, 1885 to Present," working paper, MeasuringWorth.
- Yuan, Y., 2008, "Attention and trading," working paper, University of Iowa.

**Table 1**  
**Summary statistics**

The table reports sample statistics for the media content measures used in the paper. These measures are constructed from the columns “Financial Markets” and “Topics in Wall-Street” published in the New York Times, as well as “Abreast of the market,” from the Wall Street Journal, in the period 1905–2005. We construct the “Positive” and “Negative” measures by counting the number of positive and negative words and normalizing it by the total number of words of each article, using the Loughran and McDonald (2011) dictionaries. The “Media content” variable is simply the difference between the “Positive” and “Negative” measures. All numbers are given in percentages.

	Positive words (%)			Negative words (%)			Media content		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
All articles	1.27	1.18	0.63	2.08	1.96	0.94	−0.82	−0.72	1.15
NYT articles	1.21	1.11	0.65	2.08	1.96	0.95	−0.87	−0.79	1.16
WSJ articles	1.42	1.35	0.57	2.09	1.96	0.93	−0.67	−0.54	1.12
Time-series aggregate	1.26	1.23	0.36	2.03	1.97	0.63	−0.77	−0.70	0.79

**Table 2**  
**Media content as a function of DJIA returns**

The table reports point estimates from the model

$$M_t = \beta_0 R_t + \beta \mathcal{L}(R_t) + \rho \mathcal{L}(M_t) + \eta X_t + \epsilon_t;$$

where  $M_t$  denotes the media content written between trading dates  $t$  and  $t + 1$  (for articles written after the market closed on date  $t$  by prior to opening on date  $t + 1$ ),  $R_t$  denotes the log-return on the DJIA on date  $t$  (from close on date  $t - 1$  to close on date  $t$ ). We truncate  $R_t$  at  $-3\%$  from below and  $+3\%$  from above. The set of explanatory variables  $X_t$  includes day-of-the-week dummies and a cubic function of time. The  $t$ -statistics reported in the table use Newey-West corrections with ten lags.

	Only $\mathcal{L}(R_t)$		Only $\mathcal{L}(M_t)$		$\mathcal{L}(R_t)$ and $\mathcal{L}(M_t)$	
	Estimate	$t$ -stat	Estimate	$t$ -stat	Estimate	$t$ -stat
$R_t$	0.478	91.7			0.470	95.8
$R_{t-1}$	0.149	28.5			0.077	13.6
$R_{t-2}$	0.082	15.7			0.022	3.9
$R_{t-3}$	0.073	13.9			0.022	3.8
$R_{t-4}$	0.067	12.8			0.008	1.3
$R_{t-5}$	0.054	10.3			-0.013	-2.2
$R_{t-6}$	0.040	7.6			-0.026	-4.6
$R_{t-7}$	0.049	9.3			-0.002	-0.4
$R_{t-8}$	0.041	7.9			-0.014	-2.5
$R_{t-9}$	0.045	8.7			-0.013	-2.4
$M_{t-1}$			0.217	36.0	0.141	23.3
$M_{t-2}$			0.071	11.5	0.071	11.7
$M_{t-3}$			0.058	9.4	0.054	8.8
$M_{t-4}$			0.068	11.0	0.068	11.2
$M_{t-5}$			0.056	9.1	0.072	11.8
$M_{t-6}$			0.045	7.2	0.064	10.5
$M_{t-7}$			0.023	3.8	0.029	4.8
$M_{t-8}$			0.039	6.3	0.045	7.5
$M_{t-9}$			0.036	5.9	0.043	7.1
$M_{t-10}$			0.044	7.4	0.035	6.7
adj- $R^2$		0.346		0.224		0.423

**Table 3****Media content as a non-linear function of last DJIA return**

The table reports point estimates from the model

$$M_t = f(R_t; \alpha, \beta) + \eta X_t + \epsilon_t;$$

where  $M_t$  denotes the media content written between trading dates  $t$  and  $t + 1$  (for articles written after the market closed on date  $t$  by prior to opening on date  $t + 1$ ),  $R_t$  denotes the log-return on the DJIA (truncated at  $-3\%$  from below and  $+3\%$  from above). The set of explanatory variables  $X_t$  includes 10 lags of  $R_t$  and  $M_t$ , as well as day-of-the-week dummies and a cubic function of time. The function  $f(R_t; \alpha, \beta)$  is assumed to be of the form

$$f(R_t; \alpha, \beta) = \sum_{i=1}^4 (\alpha_i + \beta_i R_t) 1_{R_t \in S_i} \quad (6)$$

where the sets  $S_i$  are:  $S_1 = (-3, -1)$ ,  $S_2 = (-1, 0)$ ,  $S_3 = (0, 1)$ , and  $S_4 = (1, 3)$ . All statistics reported in the table use Newey-West corrections with ten lags. The time period goes from January 3, 1905 through December 31, 2005, for a total of 27,448 trading days.

A. Point estimates	Intercepts		Slopes	
	$\alpha_i$	$t$ -stat	$\beta_i$	$t$ -stat
$S_1 = (-3, -1)$	-0.452	-6.1	0.203	8.6
$S_2 = (-1, 0)$	0.007	0.1	0.656	22.5
$S_3 = (0, 1)$	0.161	2.6	0.528	21.0
$S_4 = (1, 3)$	0.577	7.8	0.057	2.5
B. Tests				
	$F$ -stat	$p$ -value		
$\alpha_2 = \alpha_3$	75.1	0.000		
$\beta_2 = \beta_3$	11.3	0.001		
$\beta_1 = \beta_4$	20.0	0.000		

**Table 4**  
**Media content as a function of multiple DJIA lagged returns**

The table reports point estimates from the model

$$M_t = f_1(R_t; \alpha_1, \beta_1) + f_2(R_{t-1}; \alpha_2, \beta_2) + f_3(R_{t-2}; \alpha_3, \beta_3) + f_4(R_{t-3}; \alpha_4, \beta_4) + \eta X_t + \epsilon_t;$$

where  $M_t$  denotes the media content written between trading dates  $t$  and  $t + 1$  (for articles written after the market closed on date  $t$  by prior to opening on date  $t + 1$ ),  $R_t$  denotes the log-return on the DJIA (truncated at  $-3\%$  from below and  $+3\%$  from above). The set of explanatory variables  $X_t$  includes day-of-the-week dummies and a cubic function of time. The function  $f_j(R_t; \alpha, \beta)$  is assumed to be of the form

$$f_j(R_{t-j}; \alpha_j, \beta_j) = \sum_{i=1}^4 (\alpha_{ji} + \beta_{ji} R_{t-i}) 1_{R_{t-i} \in S_i} \quad (7)$$

where the sets  $S_i$  are:  $S_1 = (-3, -1)$ ,  $S_2 = (-1, 0)$ ,  $S_3 = (0, 1)$ , and  $S_4 = (1, 3)$ . All statistics reported in the table use Newey-West corrections with ten lags. The time period goes from January 3, 1905 through December 31, 2005, for a total of 27,448 trading days.

A. Point estimates	Intercepts		Slopes	
	$\alpha_{ji}$	$t$ -stat	$\beta_{ji}$	$t$ -stat
<b>1. Return last trading day.</b>				
$S_1 = (-3, -1)$	-0.469	-5.9	0.240	9.3
$S_2 = (-1, 0)$	-0.002	0.0	0.699	22.5
$S_3 = (0, 1)$	0.154	2.3	0.516	19.4
$S_4 = (1, 3)$	0.573	7.4	0.027	1.1
<b>2. Return two-days ago.</b>				
$S_1 = (-3, -1)$	-0.114	-1.5	0.185	7.9
$S_2 = (-1, 0)$	-0.051	-0.8	0.248	8.6
$S_3 = (0, 1)$	-0.053	-0.8	0.109	4.0
$S_4 = (1, 3)$	-0.010	-0.1	0.025	1.0
<b>3. Return three-days ago.</b>				
$S_1 = (-3, -1)$	-0.052	-0.7	0.113	4.7
$S_2 = (-1, 0)$	0.009	0.1	0.155	5.1
$S_3 = (0, 1)$	-0.006	-0.1	0.047	1.7
$S_4 = (1, 3)$	0.092	1.2	-0.057	-2.2
<b>4. Return four-days ago.</b>				
$S_1 = (-3, -1)$	-0.010	-0.1	0.094	3.7
$S_2 = (-1, 0)$	-0.003	-0.1	0.099	3.3
$S_3 = (0, 1)$	0.040	0.6	0.006	0.2
$S_4 = (1, 3)$	0.035	0.4	-0.008	-0.3

Table 4 continued.

---

B. Tests	<i>F</i> -stat	<i>p</i> -value
<b>1. Return last trading day.</b>		
$\alpha_2 = \alpha_3$	67.7	0.000
$\beta_2 = \beta_3$	20.5	0.000
$\beta_1 = \beta_4$	35.4	0.000
<b>2. Return two-days ago.</b>		
$\alpha_2 = \alpha_3$	0.1	0.903
$\beta_2 = \beta_3$	12.1	0.001
$\beta_1 = \beta_4$	21.7	0.000
<b>3. Return three-days ago.</b>		
$\alpha_2 = \alpha_3$	0.6	0.444
$\beta_2 = \beta_3$	6.9	0.009
$\beta_1 = \beta_4$	23.7	0.000
<b>4. Return four-days ago.</b>		
$\alpha_2 = \alpha_3$	4.9	0.027
$\beta_2 = \beta_3$	5.4	0.020
$\beta_1 = \beta_4$	8.0	0.005

---

**Table 5**  
**Media content response for different authors**

The table reports point estimates from the model

$$M_t = f(R_t; \alpha, \beta) + \eta X_t + \epsilon_t;$$

where  $M_t$  denotes the media content written between trading dates  $t$  and  $t + 1$  (for articles written after the market closed on date  $t$  by prior to opening on date  $t + 1$ ),  $R_t$  denotes the log-return on a given stock index (truncated at  $-3\%$  from below and  $+3\%$  from above). The regression is run independently for each of the listed ten authors of the column “Abreast-of-the-market,” published in the Wall Street Journal. The column labelled  $N$  presents the number of articles for each author. The set of explanatory variables  $X_t$  includes 5 lags of  $R_t$  and  $M_t$ , as well as day-of-the-week dummies and a cubic function of time. The function  $f(R_t; \alpha, \beta)$  is assumed to be of the form

$$f(R_t; \alpha, \beta) = \sum_{i=1}^4 (\alpha_i + \beta_i R_t) 1_{R_t \in S_i} \quad (8)$$

where the sets  $S_i$  are:  $S_1 = (-3, -1)$ ,  $S_2 = (-1, 0)$ ,  $S_3 = (0, 1)$ , and  $S_4 = (1, 3)$ . The time period is 1970–2005. All statistics reported in the table use Newey-West corrections with ten lags.

	$N$	$\beta_1$		$\beta_2$		$\beta_3$		$\beta_4$	
		Estimate	$t$ -stat	Estimate	$t$ -stat	Estimate	$t$ -stat	Estimate	$t$ -stat
Hillery	2413	0.333	3.3	0.871	9.3	0.636	6.5	0.113	1.3
O'Brien	1215	0.533	3.5	1.169	5.1	1.069	5.2	0.218	1.5
Talley	915	-0.025	-0.1	1.111	5.4	1.097	5.3	-0.258	-1.3
Marcial	625	0.899	2.3	0.430	1.5	0.830	2.8	-0.169	-0.5
Garcia	588	0.289	1.4	1.469	6.1	0.597	2.7	-0.112	-0.7
Smith	302	-0.246	-0.6	1.124	3.7	0.945	3.0	-0.078	-0.3
Wilson	251	0.037	0.1	0.988	2.9	0.656	2.4	0.397	1.3
Browning	250	0.592	1.6	0.156	0.4	0.198	0.5	-0.066	-0.2
Pettit	222	0.307	0.4	1.313	3.6	0.218	0.5	-0.587	-0.4
Sease	157	0.624	1.8	0.735	1.6	-0.017	-0.0	0.648	0.8



**Table 6**  
**Media content in different time periods**

The table reports point estimates from the model

$$M_t = f(R_t; \alpha, \beta) + \eta X_t + \epsilon_t;$$

where  $M_t$  denotes the media content written between trading dates  $t$  and  $t + 1$  (for articles written after the market closed on date  $t$  by prior to opening on date  $t + 1$ ),  $R_t$  denotes the log-return on a given stock index (truncated at  $-3\%$  from below and  $+3\%$  from above). The set of explanatory variables  $X_t$  includes 5 lags of  $R_t$  and  $M_t$ , as well as day-of-the-week dummies and a cubic function of time. The function  $f(R_t; \alpha, \beta)$  is assumed to be of the form

$$f(R_t; \alpha, \beta) = \sum_{i=1}^4 (\alpha_i + \beta_i R_t) 1_{R_t \in S_i} \quad (9)$$

where the sets  $S_i$  are:  $S_1 = (-3, -1)$ ,  $S_2 = (-1, 0)$ ,  $S_3 = (0, 1)$ , and  $S_4 = (1, 3)$ . Each column presents the point estimates for different subsamples: 1905–1930, 1931–1955, 1956–1980, 1981–2005. All statistics reported in the table use Newey-West corrections with ten lags.

---

**A. Point estimates**

	1905–1930		1931–1955		1956–1980		1981–2005	
	Estimate	<i>t</i> -stat	Estimate	<i>t</i> -stat	Estimate	<i>t</i> -stat	Estimate	<i>t</i> -stat
$\beta_1$	0.241	5.7	0.255	7.3	0.286	4.2	0.236	3.9
$\beta_2$	0.459	8.2	0.652	13.0	0.770	14.7	0.794	11.8
$\beta_3$	0.375	8.4	0.497	11.8	0.602	12.6	0.604	9.2
$\beta_4$	0.024	0.6	0.067	2.1	0.094	1.5	0.116	2.2
$\alpha_1$	-0.094	-0.8	-0.311	-2.3	-0.716	-4.7	-0.484	-2.3
$\alpha_2$	0.196	2.1	-0.050	-0.4	-0.219	-1.9	0.122	0.7
$\alpha_3$	0.319	3.4	0.030	0.3	-0.039	-0.3	0.397	2.1
$\alpha_4$	0.647	5.5	0.371	2.9	0.355	2.4	0.860	4.2

**B. Tests**

	1905–1930		1931–1955		1956–1980		1981–2005	
	<i>F</i> -stat	<i>p</i> -value	<i>F</i> -stat	<i>p</i> -value	<i>F</i> -stat	<i>p</i> -value	<i>F</i> -stat	<i>p</i> -value
$\alpha_2 = \alpha_3$	12.9	0.000	8.4	0.004	28.1	0.000	34.1	0.000
$\beta_2 = \beta_3$	1.3	0.249	5.9	0.015	5.6	0.018	4.1	0.043
$\beta_1 = \beta_4$	12.4	0.000	15.1	0.000	4.2	0.039	2.3	0.129

---

**Table 7**  
**Media content and low-frequency returns**

The table reports point estimates from the model

$$M_t = f_1(R_t; \alpha_1, \beta_1) + f_2(R_{t-4,t-1}; \alpha_2, \beta_2) + f_3(R_{t-24,t-5}; \alpha_3, \beta_3) + f_4(R_{t-252,t-25}; \alpha_4, \beta_4) + \eta X_t + \epsilon_t;$$

where  $M_t$  denotes the media content written between trading dates  $t$  and  $t + 1$  (for articles written after the market closed on date  $t$  by prior to opening on date  $t + 1$ ),  $R_t$  denotes the log-return on the DJIA (truncated at  $-3\%$  from below and  $+3\%$  from above). The variable  $R_{t,s}$  refers to the average return over dates  $t$  and  $s$ , both included. The variables  $R_{t,s}$  are expressed as daily returns, that is, they are multiplied by  $\sqrt{t - s + 1}$ , so we can keep the truncation homogeneously for all lags. The set of explanatory variables  $X_t$  includes day-of-the-week dummies and a cubic function of time. The function  $f_j(R_t; \alpha, \beta)$  is assumed to be of the form

$$f_j(R_{t-j}; \alpha_j, \beta_j) = \sum_{i=1}^4 (\alpha_{ji} + \beta_{ji} R_{t-i}) 1_{R_{t-i} \in S_i} \quad (10)$$

where the sets  $S_i$  are:  $S_1 = (-3, -1)$ ,  $S_2 = (-1, 0)$ ,  $S_3 = (0, 1)$ , and  $S_4 = (1, 3)$ . All statistics reported in the table use Newey-West corrections with ten lags. The time period goes from January 3, 1905 through December 31, 2005, for a total of 27,448 trading days.

---

**A. Point estimates**

	Last trading day		Previous trading week		Last month returns		Last year	
	$R_t$		$R_{t-4,t-1}$		$R_{t-24,t-5}$		$R_{t-252,t-25}$	
	Estimate	<i>t</i> -stat	Estimate	<i>t</i> -stat	Estimate	<i>t</i> -stat	Estimate	<i>t</i> -stat
$\beta_1$	0.217	8.3	0.170	6.7	0.027	0.9	-0.036	-1.1
$\beta_2$	0.703	22.7	0.348	11.0	0.090	2.6	0.148	3.4
$\beta_3$	0.519	19.7	0.161	5.7	0.107	3.5	0.033	0.9
$\beta_4$	0.046	1.8	0.057	2.0	0.052	1.5	0.069	2.5

**B. Tests**

	<i>F</i> -stat	<i>p</i> -value	<i>F</i> -stat	<i>p</i> -value	<i>F</i> -stat	<i>p</i> -value	<i>F</i> -stat	<i>p</i> -value
	<hr/>		<hr/>		<hr/>		<hr/>	
$\beta_2 = \beta_3$	20.7	0.000	19.1	0.000	0.1	0.714	3.7	0.053
$\beta_1 = \beta_4$	20.9	0.000	8.3	0.004	0.3	0.576	6.1	0.013

---

**Table 8**  
**Media content as a non-linear function of other indexes**

The table reports point estimates from the model

$$M_t = f(R_t; \alpha, \beta) + \eta X_t + \epsilon_t;$$

where  $M_t$  denotes the media content written between trading dates  $t$  and  $t + 1$  (for articles written after the market closed on date  $t$  by prior to opening on date  $t + 1$ ),  $R_t$  denotes the log-return on a given stock index (truncated at  $-3\%$  from below and  $+3\%$  from above). The set of explanatory variables  $X_t$  includes 5 lags of  $R_t$  and  $M_t$ , as well as day-of-the-week dummies and a cubic function of time. The function  $f(R_t; \alpha, \beta)$  is assumed to be of the form

$$f(R_t; \alpha, \beta) = \sum_{i=1}^4 (\alpha_i + \beta_i R_t) 1_{R_t \in S_i} \quad (11)$$

where the sets  $S_i$  are:  $S_1 = (-3, -1)$ ,  $S_2 = (-1, 0)$ ,  $S_3 = (0, 1)$ , and  $S_4 = (1, 3)$ . Return information is from Ken French's website. In the column labelled "VW index" we use his value-weighted index, whereas in the others we use the largest and smallest quintile portfolios in size. All statistics reported in the table use Newey-West corrections with ten lags. The time period goes from July 1st of 1963 through December 31st of 2005, for a total of 12958 trading days.

---

**A. Point estimates**

	VW index		Large stocks		Small stocks	
	Estimate	<i>t</i> -stat	Estimate	<i>t</i> -stat	Estimate	<i>t</i> -stat
$\beta_1$	0.209	4.3	0.281	5.9	0.140	2.5
$\beta_2$	0.826	18.6	0.815	18.1	0.874	17.2
$\beta_3$	0.682	16.3	0.661	15.5	0.702	15.8
$\beta_4$	0.005	0.1	0.055	1.3	0.036	0.6
$\alpha_1$	-0.580	-4.8	-0.352	-2.8	-0.692	-5.0
$\alpha_2$	-0.028	-0.3	0.067	0.7	-0.077	-0.7
$\alpha_3$	0.152	1.7	0.199	2.0	0.084	0.8
$\alpha_4$	0.685	6.0	0.658	5.5	0.558	3.9

**B. Tests**

	VW index		Large stocks		Small stocks	
	<i>F</i> -stat	<i>p</i> -value	<i>F</i> -stat	<i>p</i> -value	<i>F</i> -stat	<i>p</i> -value
$\alpha_2 = \alpha_3$	41.2	0.000	20.9	0.000	29.8	0.000
$\beta_2 = \beta_3$	5.7	0.017	6.3	0.012	6.6	0.010
$\beta_1 = \beta_4$	9.2	0.002	12.4	0.000	1.5	0.216

---

**Table 9**  
**Media content and DJIA returns along the business cycle**

The table reports point estimates from the model

$$M_t = D_t f_R(R_t; \alpha, \beta) + (1 - D_t) f_E(R_t; \alpha, \beta) + \eta X_t + \epsilon_t;$$

where  $M_t$  denotes the media content written between trading dates  $t$  and  $t + 1$  (for articles written after the market closed on date  $t$  by prior to opening on date  $t + 1$ ),  $D_t$  is a dummy variable that equals 1 if date  $t$  is in a recession (using NBER definitions),  $R_t$  denotes the log-return on the DJIA (truncated at  $-3\%$  from below and  $+3\%$  from above). The set of explanatory variables  $X_t$  includes 10 lags of  $R_t$  and  $M_t$ , as well as day-of-the-week dummies and a cubic function of time. The function  $f_k(R_t; \alpha, \beta)$  is assumed to be of the form

$$f_k(R_t; \alpha, \beta) = \sum_{i=1}^4 (\alpha_{ki} + \beta_{ki} R_t) 1_{R_t \in S_i} \quad (12)$$

where the sets  $S_i$  are:  $S_1 = (-3, -1)$ ,  $S_2 = (-1, 0)$ ,  $S_3 = (0, 1)$ , and  $S_4 = (1, 3)$ , and  $k = R, E$ . All statistics reported in the table use Newey-West corrections with ten lags. The time period goes from January 3, 1905 through December 31, 2005, for a total of 27,448 trading days.

---

**A. Point estimates**

	Expansions		Recessions	
	Estimate	<i>t</i> -stat	Estimate	<i>t</i> -stat
$\beta_{k1}$	0.236	7.0	0.242	7.4
$\beta_{k2}$	0.694	20.6	0.560	9.7
$\beta_{k3}$	0.559	19.2	0.404	7.7
$\beta_{k4}$	0.050	1.7	0.065	1.8
$\alpha_{k1}$	-0.503	-5.6	-0.206	-1.4
$\alpha_{k2}$	-0.021	-0.3	0.092	0.7
$\alpha_{k3}$	0.139	1.9	0.215	1.7
$\alpha_{k4}$	0.585	6.7	0.529	3.6

**B. Tests**

	Expansions		Recessions	
	<i>F</i> -stat	<i>p</i> -value	<i>F</i> -stat	<i>p</i> -value
$\alpha_{k2} = \alpha_{k3}$	60.8	0.000	11.2	0.001
$\beta_{k2} = \beta_{k3}$	9.6	0.002	4.0	0.046
$\beta_{k2} = \beta_{k3}$	16.7	0.000	13.0	0.000

---

**Table 10**  
**Media content as a function of last day's DJIA return, eight intervals**

The table reports point estimates from the model

$$M_t = f(R_t; \alpha, \beta) + \eta X_t + \epsilon_t;$$

where  $M_t$  denotes the media content written between trading dates  $t$  and  $t + 1$  (for articles written after the market closed on date  $t$  by prior to opening on date  $t + 1$ ),  $R_t$  denotes the log-return on the DJIA (truncated at  $-3\%$  from below and  $+3\%$  from above). The set of explanatory variables  $X_t$  includes 10 lags of  $R_t$  and  $M_t$ , as well as day-of-the-week dummies and a cubic function of time. The function  $f(R_t; \alpha, \beta)$  is assumed to be of the form

$$f(R_t; \alpha, \beta) = \sum_{i=1}^8 (\alpha_i + \beta_i R_t) 1_{R_t \in S_i} \quad (13)$$

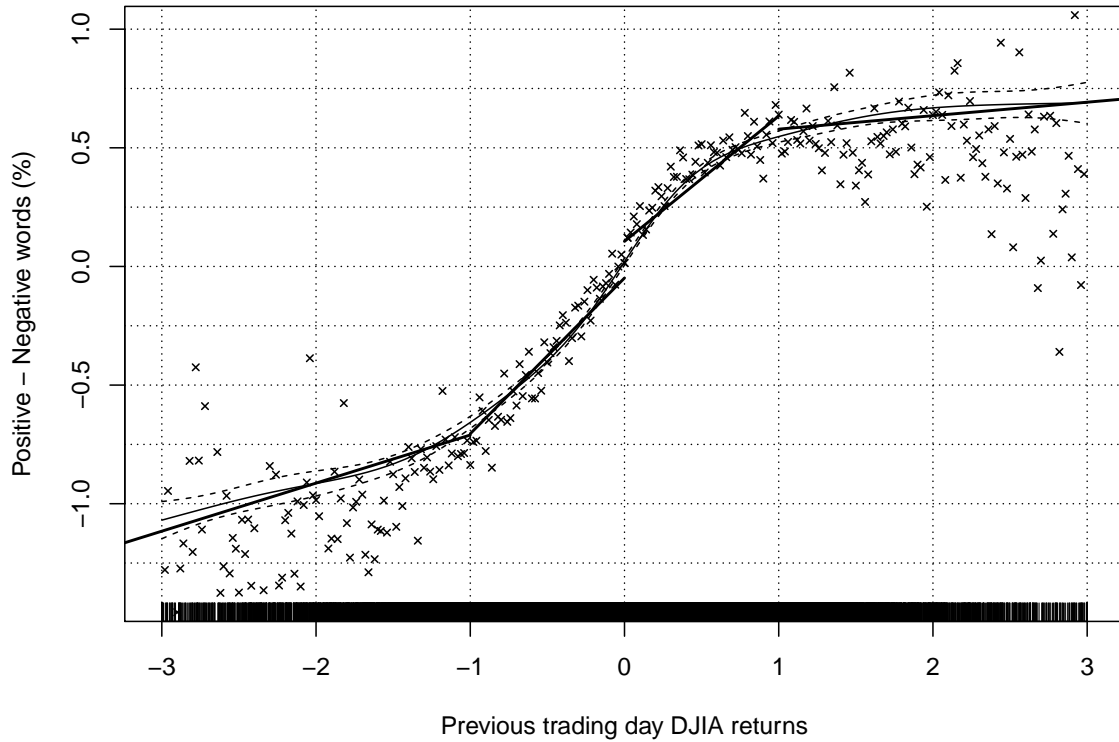
where the sets  $S_i$  are:  $S_1 = (-3, -2)$ ,  $S_2 = (-2, -1)$ ,  $S_3 = (-1, -0.5)$ ,  $S_4 = (-0.5, 0)$ ,  $S_5 = (0, 0.5)$ ,  $S_6 = (0.5, 1)$ ,  $S_7 = (1, 2)$  and  $S_8 = (2, 3)$ . All statistics reported in the table use Newey-West corrections with ten lags. The time period goes from January 3, 1905 through December 31, 2005, for a total of 27,448 trading days.

A. Point estimates	Intercepts		Slopes	
	$\alpha_i$	$t$ -stat	$\beta_i$	$t$ -stat
$S_1 = (-3, -2)$	-0.263	-1.2	0.266	3.3
$S_2 = (-2, -1)$	-0.316	-3.0	0.309	4.9
$S_3 = (-1, -0.5)$	-0.142	-1.5	0.455	4.7
$S_4 = (-0.5, 0)$	0.021	0.3	0.711	11.3
$S_5 = (0, 0.5)$	0.102	1.6	0.806	13.3
$S_6 = (0.5, 1)$	0.280	3.3	0.346	4.4
$S_7 = (1, 2)$	0.543	5.4	0.079	1.4
$S_8 = (2, 3)$	0.996	4.8	-0.100	-1.3

B. Tests	$F$ -stat	$p$ -value
$\alpha_4 = \alpha_5$	11.6	0.001
$\beta_4 = \beta_5$	1.2	0.271
$\beta_3 = \beta_6$	0.7	0.387
$\beta_2 = \beta_7$	7.5	0.006
$\beta_1 = \beta_8$	11.3	0.001

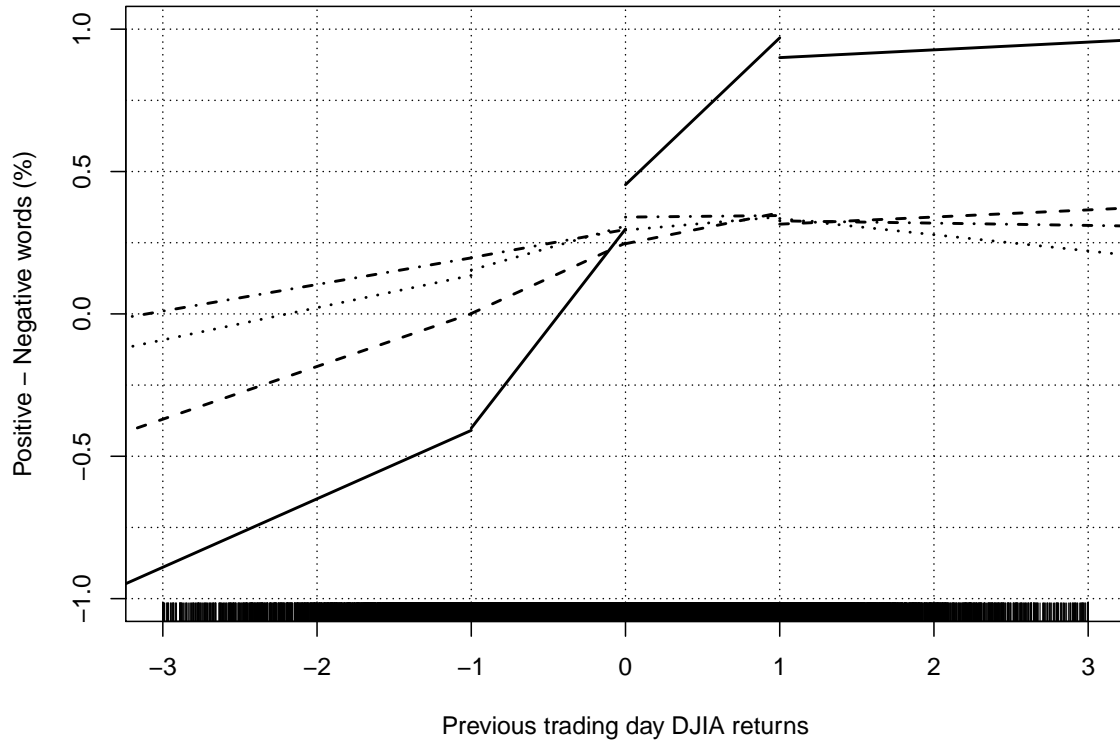
### Media content and stock returns



**Figure 1**

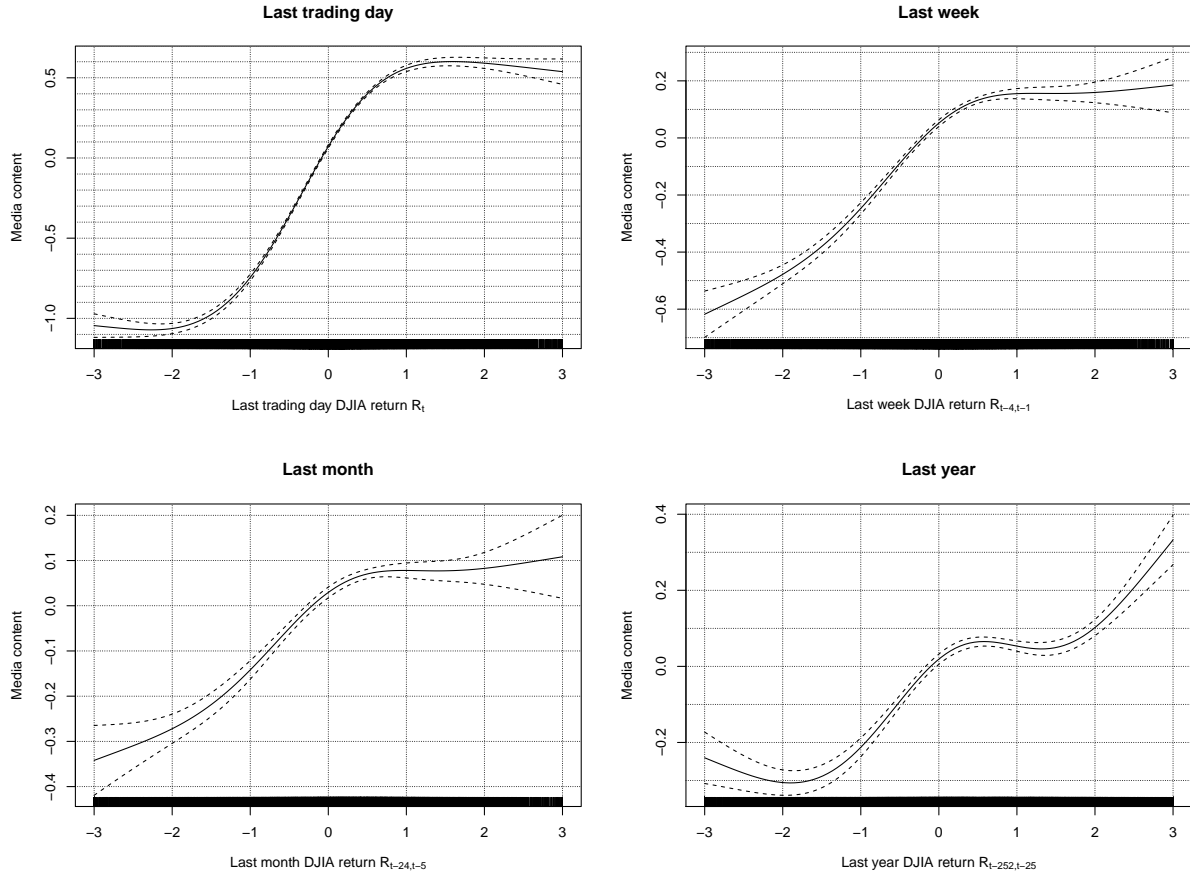
The graph presents three estimates of the relationship between media content and lagged DJIA returns. The crosses denote the average unconditional media content for intervals of 2 basis points on the range of returns  $(-3, 3)$ . The solid smooth black line presents a spline estimator, with 95% confidence intervals as dashed lines. The straight lines are OLS estimates using piece-wise linear functions over the intervals  $(-3, -1)$ ,  $(-1, 0)$ ,  $(0, 1)$  and  $(1, 3)$ . The rug in the  $x$ -axis presents the density of the right-hand side variable.

### Media content and stock returns previous four trading days



**Figure 2**

The graph presents four estimates. They measure the relationship between media content and: (1) one-day lagged DJIA returns (solid line), (2) two-day lagged DJIA returns (dashed line), (3) three-day lagged DJIA returns, and (4) four-day lagged DJIA returns. The lines in the graphs represent the OLS estimates using piece-wise linear functions over the intervals  $(-3, -1)$ ,  $(-1, 0)$ ,  $(0, 1)$  and  $(1, 3)$  of the specification in Table 4. The rug in the  $x$ -axis presents the density of the right-hand side variable.



**Figure 3**

The graph presents four estimates. They measure the relationship between media content and: (1) one-day lagged DJIA returns (top left panel), (2) lagged weekly returns, over trading days  $t - 5$  to  $t - 1$  (top right panel), (3) lagged monthly returns, over trading days  $t - 25$  to  $t - 6$  (bottom left panel), (4) lagged annual returns, over trading days  $t - 252$  to  $t - 26$  (bottom right panel). The estimates correspond to those from spline regressions with five-degrees of freedom. The rug in the  $x$ -axis presents the density of the right-hand side variable.