ISyE 2028 – Basic Statistical Methods - Fall 2015
Bonus Project: "Big" Data Analytics
Proposal (or Final Report)

The Effect of Positive Press on the Price of a Stock

by: Kristiaan Sheedy

For the final project, I tested the effect of news articles on the price movements of a stock to determine if news (aside from earnings reports) has a statistically significant effect. More specifically, I attempted to test the effect of the proportion of positive articles on a company's stock price. My initial hypothesis is that if the proportion p is within a range of .4 to .6 (40% to 60% of articles are positive), the stock price will remain within standard volatility ranges for the time period. However, if the stock's positive article proportion is outside that range, the respective stock's price will move more than 1.96 standard deviations of its typical volatility for the time period in the direction corresponding to its positive article proportion. I.e. if the proportion was .9 a stock's price would significantly increase, while a proportion lower than .4 would decrease significantly). The chosen threshold of 1.96 standard deviations of movement was chosen to demonstrate statistical significance because a stocks rate of change will be normally distributed with a standard deviation corresponding to its' volatility over time period n. As a qualifier, this problem is attempting to assess human behavior (will those who read positive articles respond by buying more of the stock etc.) and there are more factors in play than articles. As a result, any result found should be interpreted as a correlation. Linear regression modelling of the results was also conducted with the goal of establishing a correlation.

Formally, I established a null hypothesis that stocks outside the .4 to .6 positive proportion range would not move significantly (more than 1.96 STDEVs) in accordance with the aforementioned directions. My alternative hypothesis was that those outside the .4 to .6 range would in fact move more than 1.96 STDEVs in the direction that their positive proportion implied (up for greater than .6, down for less than .4). In order to test this hypothesis, I used a sample set of 9 companies from the S&P 500. I chose a sample set from this particular market subset of publically traded companies because these stocks typically receive higher news coverage and analysis and therefore the market will likely respond more quickly their news relative tot less covered stocks. Ultimately, the stocks (tickers included because a company can have more than one publically traded asset class) chosen for inclusion were AAPL, YHOO, AA, HD, MSFT, GE, BAC, F, and LUV. Stocks from several different sectors were chosen in order to provide a representative sample. Tech stocks were overrepresented in the sample based on their population within exchange traded securities, but more equitably with regard to market cap.

With regard to news articles, given limited time resources, I used a subset of popular stock publications to use to determine the proportion. These publications included Barron's, Bloomberg, the NY Times, Fortune, CNBC, CNN, and Reuter's. These sources typically provide analysis of news and business moves and are more company and asset centered than other sources which provide technical specifications and reviews of individual products. Articles from these publications will be interpreted (and recorded) as positive or not positive. Earnings reports will not be included in the articles used to

determine the proportion, although this was a moot point because none of the stocks' reported during the news collection period.

The procedure and data collection for the experiment and data collection consisted of two stages. The first stage lasted one business week. Over the course of the first stage, articles published by the aforementioned publications about companies in our sample were read and labeled as positive or not positive (not necessarily negative). Using a simple web crawler, all published articles from the publication set containing tags or strings containing the names of the aforementioned companies or their respective tickers were compiled. Articles were read and screened to assess their relevance and whether or not they were positive or not positive. Articles containing the search criteria (stock ticker or name) that were not actually about that company (a brief mention or link) were disregarded. During that week, the price information for each of the 9 stocks was also recorded. Several hundred articles were read and included in the proportion calculation.

Stocks	Total articles week	
YHOO	17	0.2941176471
AAPL	143	0
MSFT	45	0
GE	40	0.825
BAC	15	0.6666666667
AA	6	0.3333333333
F	15	0.8
LUV	5	0.2
HD	5	1

(Note, entries for AAPL and MSFT were accidentally deleted, they didn't actually have proportions of 0. Was unable to recover their row entries but I included their names and overall totals)

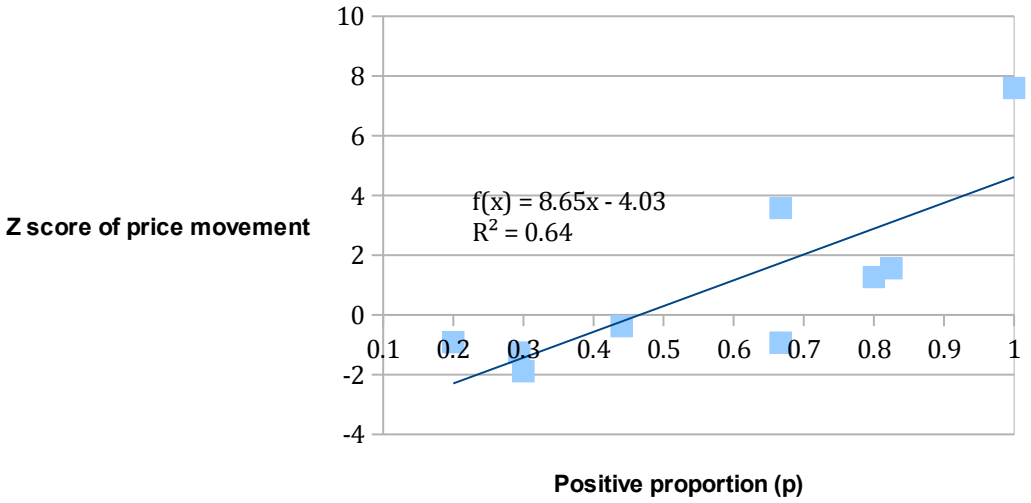
The second stage of the experiment lasted two business weeks. During the second stage, price information was recorded on a weekly basis to track price movements and to assess the impact of the first stage's proportion for each stock respectively.

After the data collection phase, the proportion data and stock movements were consolidated into excel for analysis. In order to establish a baseline for each respective stock's movement, each stocks financial data for the twelve months preceding the first stage of data collection were analyzed to determine their volatility. This data was taken from Yahoo! Finance and imported into excel. Percentage changes in price were calculated on a daily basis and used to determine standard deviations of price movement over 1 day intervals. Extrapolated from these results were the standard deviations of each respective stock's price movements for 1 week, 2 week, and 3 week intervals.

While this step established a standard deviation for each stock's movement, further steps were required to accurately normalize the data. Each stock has a beta which measures the degree to which a stock follows the S&P 500 index on average, where beta is equal to the quotient of the covariance of the stock with respect to the market over the standard deviation of the stock. These numbers were taken from Yahoo! Finance and are calculated from data over the trailing twelve months. Using these betas and

the movement of S&P 500 index over this time, each respective stock's movements in price were normalized by adjusting their movements with the movement of the index. Finally, each stock's normalized movement was converted into a z-score with their respective standard deviations. This allowed an apples to apples analysis of the sample stocks with respect to their proportions. Had the data not been normed in this way, a stock with a positive article proportion of 1 which moved 3% and one with a proportion of .6 which moved 3% would have appeared to be the same. However, if the standard deviation of the first stock was 1% over the analyzed time interval and the second had a standard deviation of 5%, the first stock would have made a statistically significant movement (3 standard deviations) while the other (moving .6 standard deviations) would not be considered significant.

Once this information was determined, several comparative measures were performed. First, an average z score of movement was taken of those stocks outside the .4 to .6 proportion range. On average, those outside of the range moved 2.44 standard deviations (directionality included). This allowed us to reject our null hypothesis and confirm the alternate with some precautions. The scale of this experiment was small. As a result, even though our sample behaved according to the alternate hypothesis, it would take a much large sample to determine if this conclusion is in fact valid. A linear regression analysis was also performed of the positive article proportion vs. the z-score of stock movement.



The results of this regression demonstrate an R^2 correlation of .64, indicating a positive correlation between the article proportion and a stock's movement. Many psychology studies (this experiment, being one mixing human behavior with quantitative movements of a stock) establish an R^2 of .4 as indicating mild correlation and a value of .6 or greater as moderate correlation. With that as a benchmark, this experiment appears to establish a moderate correlation between the proportion of recent positive articles about a stock and the movement of its market price.