THE INVARIANCE HYPOTHESIS AND NEWS

Applying the microstructure invariance hypothesis on the arrival rate of news

BAWER ARSLAN
KAJ TALLUNGS

Bachelor Thesis

Stockholm School of Economics

2020

The Invariance Hypothesis and News: Applying the Microstructure Invariance Hypothesis on the Arrival Rate of News

Abstract:

Following the methodology of Kyle et al. (2012) we investigate how the arrival rate of news articles mentioning individual stocks varies with the level of trading activity of the same securities. Defining trading activity W as the product of dollar volume and volatility, we use a sample of 243 stocks belonging to the NYSE and contrast their trading activity along with the corresponding securities' arrival rate of news articles for the same period, using news data from Factiva. Following Kyle et al. (2012), we hypothesize that the news arrival process occurs in "business time" which unfolds at a rate proportional to W to the power of two-thirds, as predicted by the market microstructure invariance hypothesis. We conclude with a dismissal of our hypothesis, as we in our regression models calculate an exponent for W considerably lower than that of two-thirds.

Keywords:

Microstructure invariance, News arrival, Trading activity, Business time, Factiva

Authors:

Bawer Arslan (23891)

Kaj Tallungs (23839)

Tutors:

Olga Obizhaeva, Assistant Professor, Department of Finance

Examiner:

Adrien d'Avernas, Assistant Professor, Department of Finance

Bachelor Thesis

Bachelor Program in Business & Economics

Stockholm School of Economics

1. Introduction

In the information age, people are witnessing an ever-increasing connectivity and access to information. This in turn raises the question of the level of synchronization between trading activity and the arrival rate of information. The relationship between the two is tested by Kyle et al. (2012) who gauged the relation between the arrival rate of news and trading activity for individual stocks. The relationship was tested through a set of different hypotheses, but most notably among their findings and most relevant for our paper was that the arrival rate of news articles into the market was found to occur at a pace nearly the same as that of the stock trading process predicted the microstructure invariance framework, put forth by Kyle and Obizhaeva (2016). The authors introduced the novel idea of microstructure invariance, which entails the study of the trading process for stocks and specifically the relationship between trading activity and business time. Business time is defined as the relative market velocity of stocks, and is described in detail later in the paper. Kyle et al. (2012) in turn expanded the scope of the microstructure invariance hypothesis by extending it from solely dealing with trading activity, to also applying the hypothesis on the process of information flow and news articles.

In this paper, we investigate how the arrival rate of news articles for a given stock varies with the same security's trading activity. Our research question is motivated by the findings of Kyle et. al (2012) where, as mentioned, it is found that the arrival rate of news articles into the market occurs at a pace near the same as that of the trading process for stocks in the microstructure invariance framework. If trading activity is defined as W, the arrival rate of news for the dataset of Kyle et al. (2012) was found to be $W^{0.68}$, whilst the stock trading process is predicted by the invariance hypothesis to be occurring in business time at a rate of $W^{2/3}$. Drawing on the close proximity of these figures, which seemingly confirms the hypothesis of Kyle et. al (2012) that information flow into the market is governed by business time, we are motivated to verify and assess the validity of the paper's results using a different dataset of stocks and news articles under a more recent timeframe, but retaining and following the authors' methodology. In other words, we will test whether business time regulates the arrival rate of news articles into the marketplace for individual stocks.

With stock data gathered from CRSP for a sample of stocks belonging to the NYSE, and news articles compiled from Factiva for the corresponding stock sample, the specific research question is: How does the arrival rate of news articles for a given stock in the NYSE vary with the same security's trading activity? The time horizon of the dataset consists of 20 months for a sample of 243 firms.

To enable empirically demonstrable hypotheses for eventual invariance in the information process, we make in the spirit of Kyle et al. (2012) two hypotheses regarding information flow and news articles, assuming that they arrive proportionally to the pace of business time. That is, for every one percentage increase in trading activity W, the arrival rate of information and news articles by extension is increased by two-thirds of one percent. If μ denotes the arrival rate of news articles, then $\mu \sim W^{\gamma}$ where $\gamma = 2/3$ per the microstructure invariance hypothesis.

Following these empirical assumptions, we perform our tests using three different distribution models for our dependent variable, the arrival rate of news. We utilize a log-linear regression to measure the relationship of our dependent variable with trading activity followed by using count-data regressions with a Poisson model and a negative binomial model respectively. In the Poisson model, our dependent variable exhibits a constant arrival rate equivalent to W^{γ} with the same variance. The negative binomial model on the other hand allows the Poisson arrival rate to vary randomly, and unlike the Poisson model allows for a variance greater than the dependent variable's mean. In all three cases, we assume market velocity $\gamma = 2/3$.

For our sample spanning 20 months throughout 2018-2020, and using our stock data from CRSP, we tested the hypothesis of the arrival rate of news equalling the pace of market velocity $\gamma = 2/3$. The tests performed utilize a log-linear, a Poisson and a negative binomial modelling of our dependant variable the arrival rate of news, where γ along with others estimates are predicted in the regression analyses. We conclude however that our hypothesis is rejected since we calculate a value for the arrival rate of news significantly lower than $\gamma = 2/3$. Closest to our hypothesis was the Poisson regression, which estimated the arrival rate of news approximately being 0.54, which is still deemed significantly lower than that of $\gamma = 2/3$.

2. Background

2.1 An introduction to microstructure invariance

Market microstructure invariance is a framework developed by Kyle and Obizhaeva (2016) culminating in the assertion of two main empirical hypotheses: The invariant distribution of both risk transfer and transaction costs for any stock across markets. These empirical hypotheses, along with the mathematical framework supporting these hypotheses are based on three transformations and definitions of the authors.

1) The first undertaking in the invariance framework is that normal stock trades are converted into bets. A bet is thought to be an original idea which in turn spurs one or several trades rooted in the same original idea, with the intention of creating idiosyncratic gain. For this reason, trading volume will be larger than bet volume, which is distilled from the total volume of trades. In this paper, trading activity is used instead of betting activity, as the latter is much more difficult to observe and consequently much more difficult to quantify.

2) Secondly, standard calendar is in this framework converted into business time, which is defined as standard calendar time between bets. Business time is therefore equal to the arrival rate of bets and can be named market velocity under this framework. Stocks which are traded often experience a higher level of market velocity, and vice versa.

3) The third and last definition underpinning microstructure invariance is that of risk, which measures the market impact of betting viewed in business time.

The framework of market microstructure invariance is such that the trading of stocks is considered a trading game in which the actors are long- and short-term players. Long-term traders trade stock with the intention of implementing a bet. This sets them apart from short-term players, such as market makers and high frequency traders who have no intention of doing so. In the market, there are countless trading games taking place concurrently. The idea behind microstructure invariance is that these trading games of bets are the same in their microstructure characteristics across all stocks when viewing them in business time, as opposed to calendar time. Business time, also called market velocity, constitutes an integral part of this paper as its relationship to information flow is studied. An important feature of business time is that it passes at a faster rate for active stocks than it does for inactive stocks, and vise versa, which highlights its relative, stock-specific nature.

The findings of the authors Kyle et al. (2016) are in part that business time proportionally varies with the betting activity of a stock to the power of two-thirds, $W^{2/3}$. The overall findings of Kyle et al. (2016) enable the prediction of a range of microstructure characteristics, such as price formations and transaction costs. Microstructure invariance deals in particular with the constant distribution of risk transfers (carried by bets) and their transaction costs. Yet again, an important distinction is that due to the difficulty of observing betting activity, this paper uses trading activity in its stead. Additionally, contrasting it with the arrival rate of news, we mimic public information flow and can thereby gauge a relationship between ideas and trades.

2.2 The mathematics of microstructure invariance

To clarify further the relationship between trading activity and the arrival rate of news which is studied in this paper, the mathematical framework behind microstructure invariance as it pertains to the relationship between business time and betting activity is demonstrated in short here. We follow the calculations of Kyle et al. (2016) where relevant for our paper, meaning aspects of the paper are omitted for either being outside the scope of this paper, or for the sake of brevity. For a more thorough mathematical explanation regarding the relationship between business time and betting activity, the reader is referenced to our Appendix. For a full inspection of the mathematics behind microstructure invariance, the reader is referenced to Kyle et al. (2016).

The mathematical framework is derived from three definitions which are mentioned earlier, namely the conversion of trades into bets, calendar time into business time and an alternative definition of risk in this framework. In essence, the authors arrive at betting activity W_B, which is the product of a bet's dollar volume and dollar volatility, being defined as:

Eq1

$$W_B = \gamma^{3/2} * E\left|I_B\right|$$

Where γ is market velocity for a given stock and $E\left|I_B\right|$ the expected risk transfer of a bet. Since the distribution of risk transfer $E\left|I_B\right|$ is hypothesised to be invariant across stocks when measured in business time, it means betting activity and market velocity have a proportional relationship. Market velocity γ is in turn decomposed into two factors, $W^{2/3}{}_B$ and $W^{1/3}{}_B$, and recalling the constant nature of risk transfer when viewed in business time, solving for γ yields:

Eq2

$$\gamma = W_B^{2/3} * E\left|I_B\right|^{-2/3}$$

Market velocity γ for a stock is therefore proportional to its betting activity to the power of two-thirds, $W_B^{2/3}$. This means that a one percent increase in the speed of business time yields a two-thirds of one percent increase in betting activity. The intuition in our paper is that information flow in the form of news mimics the arrival rate of bets, and is thought to have the same relationship to trading activity as market velocity does to betting activity.

2.3 Literature review

2.3.1 Market microstructure in general

Market microstructure invariance belongs to the wider field of market microstructure which deals with the research of financial markets. Specifically, the subject matter of microstructure research are market structures, processes and characteristics, examples of which are fair price discovery, bid-ask spread determinants and transactions costs. Following the strong development of algorithmic and electronic trading in recent decades, there has been considerable growth in the research area of market microstructure (Kissel 2013).

In a paper by Mitchell and Mulherin (1994), the authors examine the effects of public information on market activity with a dataset of daily news from Dow Jones & Company. Public information entails a number of news announcements and market activity includes trading volume and market returns. Mitchell and Mulherin (1994) conclude in their paper that there exists a direct but weak relationship between news items and market activity, and highlight the complexity in connecting market activity to observable information.

Berry and Howe (1994) studied the flow of public information into financial markets on an intraday basis using news data from Reuters' news service. The authors found the arrival rate of public information to be nonconstant, along with a positive relationship between public information flow and trading volume. Public information flow's relationship to price volatility was however deemed insignificant.

The impact of macroeconomic news on the prices of government bonds was studied by Green (2005), who finds that economic announcements have a large impact on informational trading, and that the impact of public information announcements raise the level of information asymmetry in the government bond market.

Tetlock, Saar-Tsechansky and Macskassy (2008) examined if firm earnings and stock returns could be forecasted by examining the language of firm-specific news. The authors made a set of findings, the primary one being that low firm earnings can be forecast by negative words in financial news. Additionally, negative news is incorporated into stock prices with a delay. All in all, the authors found that news stories incorporate information which is otherwise difficult to distill.

The subject of time plays a large role in the microstructure invariance hypothesis, in which it is held that stocks with a faster market velocity necessitate that more shares be traded per bet than that of slower velocity stocks. Foucault et al. (2016) studied the effect of news and the speed at which informed versus uninformed players acted on this incoming information flow. The authors concluded that the faster speculator accounted for a larger fraction of trading volume and is more correlated with short-run price volatility.

2.3.2 Market microstructure invariance in particular

Researchers in the area of market microstructure as a whole have been using different theoretical foundations as a starting point for explaining trading behaviour, for example utilising game theory or adverse selection (Kyle et al. 2012). Market microstructure invariance adds to the existing literature by bridging the existent gap between theoretical models of market microstructure and their corresponding empirical tests. In doing so, market microstructure invariance contributes significantly to the field of market microstructure by augmenting the research area's empirical applications. By modifying these common and overlapping theoretical models, the microstructure invariance hypothesis seeks to deliver accurate and testable estimates of a variety of market characteristics, such as the size of bets and their arrival rate, bid-ask spread costs and their correlation with the level of trading activity, with the results holding across all stocks, markets and across time.

The mathematical framework supporting the market microstructure invariance hypothesis is thoroughly presented in Kyle et. al (2016), and demonstrated earlier in this paper. Most relevant for our paper, business time is found to be occurring at a rate of $W^{2/3}$, assuming W is trading activity for a given stock viewed in its dollar volume and volatility. Additionally, the authors mathematically derive the invariance of both risk transfer and transaction costs from a set of definitions and equations, effectively arriving at the mathematical structure underpinning the invariance framework.

In a paper by Kyle and Obizhaeva (2016), concurrent to that of Kyle et al. (2016),the market microstructure invariance hypothesis is empirically tested on a data set of 400.000+ portfolio transitions for the US equity market, with the change in portfolios thought to be an adequate proxy for bets as novel ideas manifested in new trades. The authors' findings were in line with the predictions of the microstructure invariance framework.

3. Hypotheses

Owing to the intuitive connection between information flow and trading activity, we formulate two hypotheses following the methodology of Kyle et al. (2012) which assume the information flow process follows the spirit of microstructure invariance. This is done to extend the microstructure invariance framework's generalizations of the trading process to that of an information process. Specifically, we hypothesize an information flow and news article invariance for all assets over time.

1) Firstly, we hypothesize of an *information flow invariance* in which public and private information is assumed to arrive at a pace proportional to the rate of business time $\gamma = 2/3$ in line with the predictions of market microstructure invariance. For example, if trading activity W sees a one-percent increase, the pace of information flow will increase with two-thirds of one percent.
2) The second hypothesis of *news article invariance*, assumes that the expected arrival rate of news articles is proportional to the arrival rate of public information, which as outlined in the first hypothesis, is thought to equal $\gamma = 2/3$. To further illuminate the relationship between hypothesis #1 and #2, one can view the arrival rate of news as a proxy for the arrival rate of public information.

In summary, we hypothesize the arrival rate of news articles follows a distribution linked to the clock of business time $\gamma = 2/3$, and will be the relationship that is tested in this paper.

4. Research Design

Research design will be such that the trading activity $W_{i,t}$ for each stock i and month t is calculated using the CRSP dataset for stock returns. In line with the methodology of Kyle et al. (2012), trading activity $W_{i,t}$ is derived from the product of average daily dollar volume and volatility of stock i. Trading activity $W_{i,t}$ for a stock is in turn contrasted with the corresponding security's amount of news articles $\mu_{i,t}$ for the same time-period to estimate the relationship between the two. This estimation is primarily done with regression analysis using a negative binomial distribution for the arrival rate of news, which Kyle et al. (2012) found to be the best fit for the data used in their paper after comparing it to a log-linear model and Poisson distribution. Although we presume that the negative binomial distribution is best fitted for our data as well, we will much like Kyle et al. (2012) also compare it to the other two mentioned distributions. The motivation for these distributions lies in the nature of our news articles data set, as $\mu_{i,t}$ is a count variable assuming non-negative discrete values, rendering more common continuous distributions ineffective.

4.1 Test for the relationship between trading activity invariance and news data

The tests will be done by constructing the relationship between the arrival rate of news and trading activity through a set of equations for our data set. We know that business time runs faster for more traded stocks and runs slower for less traded stocks. The more active stock will exhibit a higher arrival rate of news articles μ per month, and the inactive stock a lower arrival rate of news μ^* per month. If business time runs H times faster for the more actively traded stock, it posits that:

Eq3 $$\mu = \mu^* * H$$

As earlier mentioned, business time is equivalent to the arrival rate of bets and is therefore more related to betting activity than it is trading activity. The difficulty in distilling independent bets from trading data however makes us instead use trading activity W as a measure of risk transfer.

Eq4 $$W = P * V * \sigma$$

If business time H is sped up, there will a two-fold and non-linear effect on its relationship to trading activity, which consists of a volume and volatility component. Firstly, trading volume will see a change proportional to the rise in business time H speed as betting volume increases. Secondly, although return variance sees an equivalent change to that of business time H and trading volume, return volatility as the square root of return variance instead sees a change proportional to $H^{1/2}$. In total, the non-linear relationship between trading activity and business time H will therefore be:

Eq5 $$W = W^{*} * H^{3/2}$$

With the expressions for the arrival rate of news and trading activities written, and their respective relationships to the clock of business time, we can observe the relationship between the arrival rate of news μ and trading activity W as:

Eq6 $$\mu = \mu^{*} * \left(\frac{W}{W^{*}}\right)^{2/3}$$

The relationship expressed in words means that for every one percent increase in trading activity, there is an expected two-thirds of one percent increase in our variable for the arrival rate of news. This is the relationship which is tested in this paper.

5. Data

As this paper seeks to investigate the relationship between the expected arrival rate of news μ and trading activity W for a sample of stocks, we use a data set of stock-specific news articles, which in turn is compared to the same securitiy's trading activity in the form of observed trading dollar volume and volatility.

5.1 News data

The data set of news is retrieved from Factiva which is a global news-related database and search engine belonging to Dow Jones & Company. The information which Factiva provides is related to one's search item, which in our case is that of a specific stock ticker. Upon search, Factiva gathers and displays the content in news, journals, magazines, transcripts among other formats related to one's query. Factive provides descriptions of news articles and items entailing their news title, date and timestamp along with either the article in its entirety or a summary of its contents. The news items in the Factiva database are not necessarily relevant to traders. The exact process through which Factiva links tickers to a news item is unknown to us. When looking at the data, any mentionings of the company name in any of the vast newswires covered by Factiva seems to generate the searchable ticker. In our data there have been obvious cases where the articles were systematically unrelated to the stock ticker we searched for. The most extreme example of this was Ethan Allen Interiors Inc. with the ticker ETH. It was an extreme outlier in our data, and upon inspection nearly all news items related to the crypto-currency Ethereum, often abbreviated as ETH.

Moreover, Factiva covers news sources from many countries and different languages, but the coverage is much smaller in languages other than English. This limits us to choose a sample of stocks from an exchange that is mostly related to news in the English language. Furthermore, the type of news data we are able to extract from Factiva is of a monthly frequency, and we could only do so by manual downloads for every individual search query due to not having access to Factiva's API. Due to the vast labor involved in manually extracting news data in such a way, we had to limit our study to a sample of a few hundred stocks of the NYSE. On top of only generating data for monthly frequencies, the timeframe for the news stretches back two years at the most. At the time of data collection, we were able to gain data from May 2018. To match this with our CRSP dataset which ends December 31st 2019, we are limited to 20 months for our time horizon.

Both the sort of news covered by Factiva and the method in which the news item is tagged with a ticker may be of notable difference from the NewsScope dataset used in Kyle et. al (2012) which provides more relevant information to draw judgement from at face value regarding the relationship between trading activity and information flow, such as view count and a sentiment indicator for the tone of the article.

5.1.1 Descriptive statistics for news data

The descriptive statistics of our dataset of news is tabulated in Table 4, presented in the appendix. We observe a total amount of observations equalling 4680, with the mean arrival rate of news for our sample of news articles being approximately 115 articles per month. The standard deviation of 260 news articles, demonstrating the occurrence of overdispersion in our dataset. The lowest value observed is zero, with the highest observed value being 3324.

5.2 Return data and sample construction

In order to generate a sample of NYSE stock we searched CRSP for all stocks at the end of our time period. This list of tickers includes common stock, ETF:s as well as other instruments. We limited this study to “ordinary common shares” (code 11) listed at NYSE. This resulted in 1248 securities.

These securities were then matched with their daily trading volume and end of day price for the full period, from which monthly dollar volume was calculated for each security. In order to gain more exposure in our sample to the larger stocks which matter more. We ranked the securities by average dollar volume over the full period, and randomly chose an equal amount of securities from each percentile group. With volume breaking points at the 10th, 20th, 30th, 40th, 50th, 60th, 80th, 90th and 95th, so that we get double exposure to the top 10%.

We calculated daily return r at day t using the end of day price P.

Eq7
$$r_t = ln\left(\frac{P_t}{P_{t-1}}\right)$$

To get average volatility σ for the month, using the number of the days in the month.

Eq8
$$\sigma = \sqrt{\Sigma(r_t^2/N)}$$

Using this together with the average dollar volume for each month we could calculate trading activity W as the product of average daily dollar volume and volatility.

5.2.1 Descriptive statistics for stock data

Our descriptive statistics are in turn derived from CRSP wherein we select data for our constructed sample of stocks, which consists of 243 firms in the NYSE, weighted with regard to their respective percentile of average dollar trading volume over the full period. The trading activity W as a product of average daily dollar volume and volatility, we observe 4860 observations for W, with a mean of $\$1224316$ and standard deviation of $\$3253490$.

6. Model Applications

Kyle et al. (2012) compared the results of a log-linear, a Poisson and a negative binomial model to one another, concluding that the negative binomial distribution was the best fitting model for the dependent variable. Other than drawing wisdom from their observations in comparing these distributions, we also find support in our data which speaks in favour of a negative binomial distribution over poisson as we experience overdispersion of our mean. This implies that a condition of the Poisson distribution, which assumes firms with the same level of trading activity also have equal means and variances. Nonetheless, we will test all three models and conclude afterwards which constitutes the best fit for our data.

6.1 Log-linear model

The log-linear model is applied by regressing the logarithm of the number of news items for a stock on the logarithm of trading activity W for the same security. Kyle et al. (2012) had too many observations where the amount of news articles was zero for the month, making the log-linear insufficient as it would mean taking the logarithm of zero. Instead, they aggregated their data into averages for different levels of trading activity. We on the other hand have not experienced this issue. The mean of our data is much greater and we only have 16 observations of zero. Therefore a log transformation can be made without much loss of data which enables us to disregard the procedure of rendering eventual zeros in the dataset non-existent. The regression expression looks as follows:

Eq9 $$ln\, N_{i,t} = \eta + \gamma * ln\left(\frac{W_{i,t}}{W^*}\right) + \varepsilon_{i,t}$$

Where $N_{t,i}$ denotes the distribution of the news articles. The equation introduces a benchmark stock which functions as a scaling constant, with trading activity $W^* = (40) * (10^6) * (0.02)$ where price being $40 per share, one million shares traded per day and a daily volatility of 2%. This benchmark stock is intentionally the same as Kyle et al. (2012). The benchmark stocks average number news articles per month μ^*, is defined as $\eta = ln(\mu^*)$. Developing the expression to solve for the arrival rate of news per month $N_{i,t}$ gives us:

Eq10 $$N_{i,t} = e^{\eta + \gamma * ln\left(\frac{W_{i,t}}{W^*}\right)} \cdot Z_{t,i}$$

Where $Z_{t,i}$ denotes a random variable whose mean equals one. If the random random variable has a variance equalling zero, $Z_{t,i}$ assumes a constant nature equalling one. Furthermore, the new expression means that the benchmark stock's average arrival rate of news per month is calculated from e^{η}. As we are testing the invariance hypothesis of market microstructure, market velocity γ assumes a value of two-thirds.

6.2 Poisson model

A Poisson distribution suits the modelling for count data well, which our dependent variable $\mu_{i,t}$ constitutes. The Poisson model assumes however that there is no overdispersion of the mean for our dependent variable which runs contrary to our data, and could consequently pose an issue, rendering the negative binomial model as more fitting. If the total distribution of the amount of news articles is denoted $N_{t,i}$ for stock i at month t, the Poisson model yields the following probability mass function

Eq11 $$f\left(N_{i,t} \mid W_{i,t}\right) = \frac{e^{-\mu(W_{i,t})} * \mu\left(W_{i,t}\right)^{N_{i,t}}}{N_{i,t}!}$$

In turn, the expected number of news items $\mu_{i,t}$ is calculated as a function of trading activity $W_{i,t}$:

Eq12 $$\mu\left(W_{t.i}\right) = e^{\eta + \gamma * ln(\frac{W}{W^*})}$$

As with the log-linear model, the term e^{η} entails the benchmark stock's average number of news articles per month, with W^* constituting its trading activity. As mentioned, the Poisson model assumes that the variations of our dependent variable $\mu\left(W_{t.i}\right)$ takes place within the Poisson distribution, rendering other explanatory variables for its variance irrelevant.

6.3 Negative binomial model

Although based on a Poisson arrival rate for its dependent variable, a negative binomial modelling allows the news items arrival rate to vary randomly unlike the Poisson model. The great variance found in the data set is modeled as a Gamma distribution with a variance of α.

Eq13 $$\mu(W_{t.i}) = e^{\eta + \gamma * ln(\frac{W}{W^*})} * G_{t,i}(\alpha)$$

The Gamma variable G is modelled to have a mean equal to one, and a variance equal to α. Additionally, the constant term e^{η} denotes the average number of news items reported per month for our benchmark stock, with market velocity $\gamma = 2/3$ in accordance with the prediction of the invariance framework. Using this modelling framework, the arrival rate of news in a month can be observed to vary for three different reasons. Firstly, our dependent variable, modelled as a Poisson arrival rate, varies in part due to different levels of trading activity for different stocks. Secondly, the inclusion of the Gamma distribution captures variation else unaccounted for by the Poisson distribution. Thirdly, assuming a certain Poisson arrival rate, there is also random variation in the number of actual Poisson events which occur.

7. Results

The results of our model applications are discussed here, beginning with the log-linear results and afterwards proceeding with the Poisson and negative binomial models respectively.

7.1 Log-linear model

The results from our log-linear model are presented below in Table 1. We report a relational value between trading activity and the arrival rate of news as approximately $\gamma = 0.39$. Additionally, its standard deviation is calculated as $\sigma = 0.07$. As we hypothesised of the trading process unfolding at $\gamma = 2/3$ which the invariance hypothesis predicts, we consequently dismiss the hypothesis as holding for our dataset of news gathered from Factiva and our sample of stock belonging to the NYSE. The dismissal comes in part from intuitive sense as the figures are relatively wide apart, and additionally as is stated in Table 1 the hypothesis is statistically rejected at a p-value of zero. For our intercept, we find that it approximately equals $\eta = 4.22$ with a standard deviation of roughly $\sigma = 0.17$. The coefficient of determination equals $R^2 = 0.3952$.

Table 1

We observe below the results from our log-linear model with OLS estimates. These estimates are from the following regression:

$$ln\, N_{i,t} = \eta + \gamma * ln\left(\frac{W_{i,t}}{W^*}\right) + \varepsilon_{i,t}$$

Where $N_{i,t}$ denotes the average number of news items for stock i per month t from our sample. Each observation corresponds to a specific value for both stock i and month t. The variable $W_{i,t}$ specifies the trading activity for a specific stock at a certain month, and W^* the trading activity for our benchmark stock. It functions as a scaling constant containing the benchmark stock's share price at $40, with one million shares traded per day and a daily volatility of 2% equalling, assuming a total

Table 1: OLS Estimates for The Number of News Articles

	Log-linear model
γ	0.3939 (0.007)
η	4.2239 (0.1717)
Hypothesis of Trading activity invariance: γ = 2/3	
p-value	0
R-squared	0.3952
#Observations x #Months	4844

value of $W^* = 40 * 10^6 * 0.02$. *Standard error values are found in parenthesis.*

7.2 Count-Data Models

7.2.1 Poisson model

Starting with the Poisson model, we can observe in Table 2 that the relation between our dependent variable and trading activity for our sample of stocks is estimated at approximately $\gamma = 0.54$ with a standard deviation of $\sigma = 0.008$ which is considerably low. The value of the slope with the Poisson model consequently puts us closer to the predictions of the invariance framework, but the distance remains nonetheless quite significant from $\gamma = 2/3$. This is not surprising considering that the Poisson model is a count-variable model, unlike the log-linear model, and that we are dealing with count data in this paper. The intercept for our regression is valued at $\eta = 4.85$ with a corresponding standard deviation of $\sigma = 0.0014$. Moreover, as stated in Table 2, the p-value for our hypothesis when running a Poisson regression confirms the dismissal of our hypothesis that $\gamma = 2/3$. The $Pseudo\ R^2$ score for the test is $Pseudo\ R^2 = 0.4239$.

7.2.2 Negative binomial model

Moving on to the negative binomial model, we report a slope of $\gamma = 0.54$ and a standard deviation of $\sigma = 0.008$ which, much like the standard deviation of the Poisson model, is low. Our intercept is estimated to be approximately $\eta = 4.78$ with a standard deviation of roughly $\sigma = 0.016$. As the negative binomial model introduces a Gamma variable with the mean α, we observe this value as being $\alpha = 1.0764$ with a standard deviation of approximately $\sigma = 0.0198$. A condition of the Poisson model is that $\alpha = 0$, which is not fulfilled by our data as we observe overdispersion of our variance for our dependent variable, which confirms the motivation for choosing a negative binomial model since it allows for overdispersion. Despite this however we experience a significantly lower value for $Pseudo\ R^2 = 0.0469$, which gives reason to doubt the value's implications at face value as it seems abnormally low in comparison with the $Pseudo\ R^2$ of the Poisson model.

Table 2

We observe below the estimates of our count regression models with the results of the Poisson model on the left side, and the results of the negative binomial model on the right side.

The Poisson regression models the arrival rate of news $\mu_{i,t}$ for stock i and month t as:

$$\mu_{i,t} = e^{\eta + \gamma * ln(\frac{W}{W^*})}$$

And for the negative binomial model, the regression models the arrival rate of news $\mu_{i,t}$ for stock i and month t as:

$$\mu = e^{\eta + \gamma * ln(\frac{W}{W^*})} * G_{t,i}(\alpha)$$

Where $G_{t,i}$ is a Gamma variable with a mean equalling one and a variance equal to α. Each observation corresponds to a specific value for both stock i and month t. The variable $W_{i,t}$ specifies the trading activity for a specific stock at a certain month as a product of average daily dollar volume and dollar volatility. W^* the trading activity for our benchmark stock. It functions as a scaling constant containing the benchmark stock's share price at $40, with one million shares traded per day and a daily volatility of 2% equalling, assuming a total value of $W^* = 40 * 10^6 * 0.02$. Standard error values are found in parenthesis.

Table 2: Count Regression Estimates for The Number of News Articles

	Poisson model	Negative Binomial Model
γ	0.5432 (0.008)	0.3419 (0.0054)
η	4.8521 (0.0014)	4.7834 (0.0161)
α		1.0764 (0.0198)
Hypothesis of Trading activity invariance: $\gamma = 2/3$		
p-val	0	0
Pseudo R-squared	0.4239	0.0469
# Observations x # Months	4860	4860

7.3 Comparing and discussing results

The results provide some mixed interpretations. An important distinction is that when computing our log-linear model, we forgo the exact methodology of Kyle et al. (2012), as the authors compile and aggregate their data to avoid taking the logarithm zero stemming from many stocks not having any news report about them for some time periods. This issue is nearly irrelevant to us, as we have few instances overall where the arrival rate of news equals zero for any given firm in our sample. With this in mind, we construct a model which slightly differs from that of Kyle et al. (2012), and in turn most likely contributes to a different result. The log-linear model yields a value of 0.39 for the arrival rate of news, which differs highly from $\gamma = 2/3$. Although the log-linear model might not be optimal in treating count variables, it delivers a result relatively similar to those of our count-data models. Additionally, the R-squared value yielded by the log-linear model equals $R^2 = 0.3952$.

Furthermore, perhaps the most peculiar finding belongs to that of our Poisson regression. Although the dataset exhibits overdispersion, which stands in contrast to a condition of the Poisson model, the Poisson regression provides us with a Pseudo R-squared value of $Pseudo\ R^2 = 0.4239$. Worthy of mention is that this metric differs from that of the R-squared value in the log-linear model, and could make the estimates unfit for comparison. Otherwise, the Poisson regression estimates a value for the arrival rate of news approximately equal to 0.54, and is the value closest to our hypothesis of it equalling $\gamma = 2/3$. However, the value remains significantly lower than our hypothesis of it equalling two-thirds.

In the negative binomial regression, overdispersion is taken into account and the negative binomial model includes a Gamma variable accounting for unexplained variance α. In our regression analysis, we find that $\alpha = 1.0764$, which serves as a motivator to dismiss the Poisson model, which assumes $\alpha = 0$. Despite this fact, we generate results with the negative binomial model that are further away from our initial hypothesis $\gamma = 2/3$ in addition to a significantly lower Pseudo R-squared value. The negative binomial regression estimated the arrival rate of news at a value of 0.3419 and Pseudo R-squared as approximately $Pseudo\ R^2 = 0.047$. The reason for the large contrast between the values generated by the Poisson and negative binomial regressions respectively proves a source of uncertainty when interpreting their meaning. Most reasonably one would assume the large differences have arisen from a form of miscalculation in the statistical procedures of these values, assuming the dataset is not the issue. We have difficulty otherwise explaining why a Poisson model performs better as an explanatory model than a negative binomial model when dealing with a count variable exhibiting overdispersion, and in our case a rather large overdispersion.

Moreover, we find it noteworthy to discuss the importance of our article data and contrast it with the news data of Kyle et al. (2012). Undoubtedly, varying types of news data can provide different explanatory values which affect the end-result in determining eventual relationships between variables. This paper uses news data gathered from Factiva, and constitutes article data rather different from that of the data used in Kyle et al. (2012), who used data from Thomson-Reuters provided by NewsScope. Whilst Factiva serves as a form of news article database, with rather rudimentary filtering options, NewsScope appears much more geared towards drawing judgement of the compiled news' potential effect on the reader and trading activity. NewsScope provides a range of descriptive factors for the news items it provides, for example how substantive the news item might be for the company in question, the tone of the article and how many views the item has had. In this sense, it appears that NewsScope is a more refined news source and by extension better fitted as a data source for the development of meaningful statistical relationships than Factiva.

The exponent γ being assuming a lower value in our estimates than it in Kyle et al. (2012) means in our case that a relative change in W gives a smaller relative change in the arrival rate of articles, compared to our benchmark stock. This could perhaps be explained by the news data in Factiva being less related to betting speed than what the Thompson-Reuters dataset is, especially as Factiva is a broad news aggregation database. It posits that the kind of news which Factive provides could have less of an economic relevance than Thompson-Reuters, yielding a lower gamma value.

Additionally, the hypothesis for a link between the arrival rate of news and betting frequency is that each bet has a set cost assigned to the entity making the bet, alongside each article having a set cost for its production. On average this betting cost, or the production cost of the article, should be regained through the economic value of bringing this information to the market. If news articles addressing a specific security get tagged for a reason other than its relevance to the security's valuation, it would negatively impact our assumption that each article on average should have the same economic value, which could make the news data less valuable as an explanatory variable.

We therefore believe that it is likely that different compositions, in what type of newswires are aggregated, should impact the gamma value of the data. In Kyle et al. (2012) the gamma value varied slightly over time, and made a small jump due to what was speculated by Kyle et al. (2012) to be a change in composition.

8. Conclusion

Using news data related to a sample of stocks from Factiva, we examine how information flow, proxied by the arrival of news, relates to a stock's trading activity for our sample of 243 stocks across 20 months. Running empirical tests we model our arrival rate of news using three different models; a log-linear, a Poisson and a negative binomial model. We hypothesize that information flow and by extension the arrival rate of news will mimic the pace of business time as specified by the microstructure invariance hypothesis. Specifically, the hypothesis is that for every one percent increase in trading activity, the arrival rate of news will speed by two-thirds of one percent. Having regression analyses for the three specified models, we conclude there is no empirical support for our hypothesis in this paper as we generate values significantly below that of market microstructure invariance's prediction. With only one kind of news data, news articles, and using only one news source, Factiva, we highlight the fact that different news sources might yield a different result from that of ours, along with other sources of information either from, or of relevance to, financial markets.

9. References

Published papers

Berry, T. D., & Howe, K. M., 1994, Public information arrival, The Journal of Finance, 49(4), 1331-1346.

Foucault, T., Hombert, J., & Roşu, I., 2016, News trading and speed, The Journal of Finance 71(1), 335-382.

Green, T. C., 2004, Economic news and the impact of trading on bond prices, The Journal of Finance 59(3), 1201-1233.

Kyle, A. S., Obizhaeva, A. A., & Kritzman, M., 2016, A Practitioner's guide to market microstructure invariance, The Journal of Portfolio Management, 43(1), 43-53.

Kyle, A. S., & Obizhaeva, A. A., 2016, Market microstructure invariance: Empirical hypotheses, Econometrica 84(4), 1345-1404.

Kyle, Albert S., et al., 2012, News articles and the invariance hypothesis, AFA 2012 Chicago Meetings Paper.

Mitchell, M. L., & Mulherin, J. H., 1994, The impact of public information on the stock market, The Journal of Finance 49(3), 923-950.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S., 2008, More than words: Quantifying language to measure firms' fundamentals, The Journal of Finance 63(3), 1437-1467.

Monographs

Kissell, Robert L., 2013, The science of algorithmic trading and portfolio management, page 47, (Academic Press).

10. Appendix

10.1 The mathematics of business time

Before expressing the relationship between betting activity and market velocity, we must first outline some prerequisite expressions following the calculations of Kyle et. al (2016). Betting activity is the product of a bet's dollar volume and dollar volatility. This means we begin with calculating betting volume V_B followed by betting volatility σ_B. Beginning with betting volume, which is defined as:

Eq14 $$V_B = \gamma * E|Q|$$

Where γ is market velocity for a given stock, also dubbed business time or the arrival rate of a bet, and $E|Q|$ is the expected number of shares of a stock traded by an investor in one bet.

Next up is betting volatility σ_B which is derived from stock return volatility σ_R adjusted for a volatility multiplier Ψ. Betting volatility is expressed by:

Eq15 $$\sigma_B = \Psi * \sigma_R$$

Following the definitions of betting volume and volatility, one can express the function for risk transfer I_B carried by a bet, as:

Eq16 $$I_B = P * Q * \sigma_B * \gamma^{-1/2}$$

Where P equals the price of a share of stock at bet implementation time and Q equals the amount of shares traded in a bet. For microstructure invariance to hold, these values must be scaled by business time, which is done with the negative root of velocity for the stock. Assuming P and σ_B are held constant, the equation necessitates that a higher velocity γ for a stock will imply a larger number of shares traded per bet, Q. This condition explains why it is that business time passes at a faster right for highly traded stocks, and vice versa.

Equations 14 to 16 have thus far set the stage for expressing betting activity W_B as a product of dollar volume and dollar volatility of bets. Betting activity W_B is defined as:

Eq17 $$W_B = P * V_B * \sigma_B$$

Then, we rewrite equation 16 which defines risk transfer I_B to solve for Q.

[Eq16] $I_B = P * Q * \sigma_B * \gamma^{-1/2}$ is rearranged to...

Eq18 $E|Q| = \sigma_B^{-1} * \gamma^{1/2} * P^{-1} * E\left|I_B\right|$

Having this expression for $E|Q|$ allows use a different expression for equation 13: $V_B = \gamma * E|Q| = (\gamma) * (\sigma_B^{-1} * \gamma^{1/2} * P^{-1} * E\left|I_B\right|)$, which is in turn imposed into the expression for W_B in equation 17. Following a set of simplifications, one arrives at...

Eq19 $W_B = \gamma^{3/2} * E\left|I_B\right|$

The velocity factor $\gamma^{3/2}$ in equation 17 can be decomposed into two different components; these are $W^{2/3}{}_B$ and $W^{1/3}{}_B$. Recall the hypothesis of risk transfer I_B being invariant across stocks, which effectively means betting activity only varies with the market velocity of a security, and not the bet's risk transfer.

The variation in market velocity is thus proportional to...

Eq20 $\gamma = W_B^{2/3} * E\left|I_B\right|^{-2/3}$

This means that velocity γ for a stock is proportional to its betting activity to the power of two-thirds, $W_B^{2/3}$, since the distribution of risk transfer $E\left|I_B\right|$ is invariant across stocks when measured in business time. In layman's terms, this means that a one percent increase in market velocity causes a two-thirds of one percent increase in betting activity.

10.2 Descriptive statistics

Below in Table 3 and Table 4 respectively, we present the descriptive statistics for our news articles and trading activity W. Trading activity W is defined as the product of average daily dollar volume and volatility, and contains data points for our sample of 243 stocks spanning for 20 months.

Table 3: *News articles*

Below we observe in the first column our variable articles, which denotes the arrival rate of news on a monthly basis for our sample of stocks spanning 20 months throughout 2018-20. The second column describes the total amount of observations, with the third column presenting the mean arrival rate of news per month for our dataset. The fourth column recounts the standard deviation for our dataset, with the fifth and sixth columns presenting the lowest and highest respective values for our variable.

Table 3

. summarize articles

Variable	Obs	Mean	Std. Dev.	Min	Max
articles	4,860	114.7751	260.1983	0	3324

Table 4: *Trading activity*

Below we observe in the first column our variable articles, which denotes the arrival rate of news on a monthly basis for our sample of stocks spanning 20 months throughout 2018-20. The second column describes the total amount of observations, with the third column presenting the mean arrival rate of news per month for our dataset. The fourth column recounts the standard deviation for our dataset, with the fifth and sixth columns presenting the lowest and highest respective values for our variable.

Table 4

. summarize w

Variable	Obs	Mean	Std. Dev.	Min	Max
w	4,860	1224316	3253490	.000178	6.30e+07