

MAXIMUM LIKELIHOOD ESTIMATION OF A POISSONIAN COUNT RATE FUNCTION FOR THE FOLLOWERS OF A TWITTER ACCOUNT MAKING DIRECTIONAL FORECASTS OF THE STOCK MARKET

GRAHAM L. GILLER

ABSTRACT. We derive expressions of use in the maximum likelihood estimation of a parameterized growth rate where the quantity growing is a *Poissonian* count rate parameterized in such a manner as to make it suitable to measure the number of *Twitter* accounts following an account that makes directional forecasts of the stock market. We use these expressions to estimate the model for data collected for a forecasting system publicised during the Spring of 2009. We find a positive correlation between success and an increase in the number of followers, and use the maximum likelihood ratio test to reject the null hypothesis (of no correlation) with a confidence of better than 95%.

1. INTRODUCTION

*Twitter*¹ is an internet based *social networking* business. It permits any user, free of charge, to broadcast short (140 character) messages (which are known as *tweets*) to the entire subscriber base (known as the *public timeline*) via various client programs. In normal usage, however, subscribers do not listen to this public timeline, but rather to a subset of these messages that come from a set of users that the subscriber has elected to *follow*.

In April, 2009, Giller Investments began an experiment with *Twitter* to investigate the usage of social networking within the context of the commodity trading advisor business[1]. The system has been used to broadcast a trade blotter to any users who wish to subscribe to it. The data sent represents an historical record of trades executed by our firm for a model account and is not customized for or targeted to any individual user. The trade record is published at the start of the day, the trades are held throughout the day, and closed at the end of the trading day. Therefore, the success of the prediction is essentially contemporaneous information to a subscribers decision to follow the account posting the trade data.

2. THE GROWTH RATE PROCESS

Let $\{n_t\}_{t=1}^T \in \mathbb{Z}^+$ be a sequence of random counts drawn from a Poisson distribution[2] with a growing expectation. i.e.

$$(1) \quad \Pr(n_t = N | \mu_t) = \frac{\mu_t^N e^{-\mu_t}}{N!}.$$

Let μ_t be the expected number of followers on a given day, t . For constant absolute growth with have

$$(2) \quad \mu_t = (g + h s_t) t.$$

Date: June 23, 2009. *Giller Investments Research Note: 20090618/1.*

Graham Giller is President and CIO of Giller Investments (New Jersey), LLC. *blog@gillerinvestments.com.*

¹<http://twitter.com>

where g is the growth rate, $s_t \in \{-1, 1\}$ is a *success metric* for directional forecasts² and h measures the influence of forecasting success on the period specific growth rate. We model μ_t as the expected number of followers measured at the end of the business day for which potential followers were able to assess the success s_t for a forecast made at the start of that day. In proposing this model we assume that there is *no serial correlation* in the fluctuations of the actual follower counts n_t about the growing mean μ_t . We chose constant absolute growth as this model seems more likely to apply for the *Twitter* paradigm than the alternative simplistic null hypothesis, which is constant relative growth.

3. THE MAXIMUM LIKELIHOOD ESTIMATORS

With the assumption of zero serial correlation, we may write the log-likelihood as

$$(3) \quad \mathcal{L}(g, h | \{n_t, s_t\}_{t=1}^T) = \sum_{t=1}^T n_t \ln(g + hs_t) - \sum_{t=1}^T (g + hs_t) t + \sum_{t=1}^T n_t \ln t - \sum_{t=1}^T \ln n_t!$$

Solving for the roots of the the partial derivatives of Equation 3 w.r.t. g and h gives the estimates \hat{g} and \hat{h} , respectively. These are the solutions to

$$(4) \quad \sum_{t=1}^T \frac{n_t}{g + hs_t} = \frac{1}{2} T (T + 1)$$

$$(5) \quad \text{and } \sum_{t=1}^T \frac{n_t s_t}{g + hs_t} = \sum_{t=1}^T s_t t.$$

At this point we could simply apply numerical optimization to Equation 3 to obtain the values of the estimators directly. However, let us assume an almost efficient market. i.e. That the forecast indicators are only marginally successful and that the followers are sufficiently rational to realize this. Specifically this requires that the influence of the success metric be small

$$(6) \quad g \gg hs_t \forall t;$$

and, that the average success be close to zero, i.e.

$$(7) \quad 0 \leq \frac{1}{T} \sum_{t=1}^T s_t \ll 1.$$

We may use Equation 7 to simplify Equation 4. We know that all of the terms in Equation 4 are large except for hs_t , so we eliminate that term to give

$$(8) \quad \hat{g} \simeq \frac{2}{T(T+1)} \sum_{t=1}^T n_t = \frac{2\bar{n}}{T+1}.$$

Note that this expression is approximately twice the average number of followers divided by the length of time the experiment has been run. For a deterministic growth, it is simple to show that this expression is exactly correct.

²i.e. s_t is an indicator variable that takes the values 1, when the forecast of direction is correct; -1 , when the forecast of direction is incorrect; we treat the indicator as *undefined* on days when no forecast was made.

Using the assumption of Equation 6 allows us to apply the binomial theorem to reduce the denominator in Equation 5. To first order, we have

$$(9) \quad \sum_{t=1}^T \frac{n_t s_t}{g} \left(1 - \frac{h s_t}{g} \right) \simeq \sum_{t=1}^T s_t t.$$

Rearranging this expression, and using $s_t^2 = 1$, gives

$$(10) \quad \hat{h} = \frac{2}{T(T+1)} \sum_{t=1}^T s_t (n_t - \hat{g}t).$$

With our assumption of weak forecasting power, from Equation 7, we see that this expression is approximately twice the covariance between the forecasting success and the excess growth, divided by the length of time the experiment has been running.

The asymptotic maximum likelihood estimators developed in this section have the advantage of simplicity and of *making sense* as reasonable formulae. They also avoid the calculation of logarithms, which can be computationally expensive. However, these are point estimates and, for a proper statistical analysis, it is preferable to work with interval estimates, which may be derived from the full likelihood function. Therefore, in actual data analysis, we propose to use these asymptotic point estimates as the starting points for a traditional numerical analysis.

4. DATA ANALYSIS

Giller Investments (New Jersey), LLC has published to Twitter records of directional trades done on equity index contracts listed on the the CME Group's GLOBEX platform. Specifically, the trades have been directional trades in the *Dow Jones Industrial Average* and *NASDAQ-100* e-mini futures contracts, listed under the tickers *YM* and *NQ* respectively. Futures trades are entered approximately 30 minutes after market open and positions held during the day, prior to liquidation immediately before the close³. The publication to Twitter commenced in early April, 2009, and is ongoing. Due to the manner in which the trade data is published, the Twitter subscriber audience is fully aware of the success of each days trading by the end of the day. Some examples of the message formats used in this experiment are given in Table 1.

We take as our data the measured number of distinct Twitter followers on a given day⁴. For the success metric we take the sign of the net profit for the trades publicized on that day. The dataset contains a total of 45 usable observations, from 04/09/2009 to 06/22/2009 inclusive. The data is tabulated in Table 2. Note that this analysis is independent of the global success of the trading algorithm, which we do not analyze here; our sole concern is to whether Twitter subscribers are inclined to follow a successful trader.

Using the asymptotic estimators of Section 3, we find $\hat{g} = 0.89$ and $\hat{h} = 0.07$ (both in units of *followers per trading day*). We use a numerical maximization procedure⁵ to obtain interval estimates of $\hat{g} = 0.7927 \pm 0.0298$ and $\hat{h} = 0.0112 \pm 0.0299$. The maximum

³The trades themselves are based on a proprietary forecasting and trading model, but the precise for of this system is not relevant to the discussion here

⁴We use the count of distinct followers because some subscribers repeatedly follow and then *unfollow* an account in an attempt to induce the owner of that account to follow the following subscriber. This strategy exists because the account being followed is notified by email when another subscriber elects to follow it. It is viewed as Twitter *good manners* to follow someone who elects to follow you, making them your *friend*.

⁵The time-series analysis program, RATS, published by Estima, Inc. of IL. The computation is performed by the MAXIMIZE command in v7.2 of the program[3].

likelihood ratio test can be used to test the hypothesis that $h \neq 0$ versus the null hypothesis that success has no influence on the number of subscribers ($h = 0$). We find a χ_1^2 statistic of 4.25 with a p -Value of 0.039, indicating that we may reject the null hypothesis with a confidence of better than 95%.

5. CONCLUSIONS

We have examined a small dataset for an experiment in the use of the social networking site Twitter to publicize a record of directional intraday index futures trades. We find that we can accept the hypothesis that the number of followers is influenced by the contemporaneous success of each days trading with a confidence of better than 95%. Although, this is a statistically weak result, it is derived from a meagre sample of only 45 trading days, and as such it warrants further data collection. The methodology developed here is framed in a manner which makes it's use suitable for the publication of any information that may be determined to be either right or wrong, and as such it is not actually restricted to the analysis of forecasts of the stock market. This method could equally be applied to the analysis of any binary outcome, such as a sporting event or even weather forecasting.

REFERENCES

- [1] Giller, G.L., "Index Futures Trades on Twitter — A Web 2.0 Experiment Part II," <http://blog.gillerinvestments.com/post/2009/04/20/Index-Futures-Trades-on-Twitter-A-Web-20-Experiment-Part-II.aspx>, 2009.
- [2] Evans, M., Hastings, N., & Peacock, B., "Statistical Distributions," 3rd. Edn., pp. 155–160, John Wiley & Sons, Inc., 2000.
- [3] Doan, T.A., "RATS Version 7 User Guide," pp. 286–287 Estima, 2007.

TABLES

Type	Name	Example Message
I	VWAP Summary (Closed Trade)	06/22/2009 \$NQU #SHORT SLD 1442.23 BOT 1426.64, GAIN 15.58; \$YMU #SHORT SLD 8368.64 BOT 8296.94, GAIN 71.70; http://is.gd/tGyI #futures #emini
I	VWAP Summary (Open Trade)	06/22/2009 \$NQU #SHORT SLD 1451.75, LOSS 1.75; \$YMU #SHORT SLD 8403.00, LOSS 2.00; http://is.gd/tGyI #futures #emini
II	Trade Record (Buy Trade)	16:14:47 BOT 9 \$NQU 1427.5 Data delayed. In real time at http://is.gd/tGyI #futures #emini
II	Trade Record (Sell Trade)	08:56:14 SLD 2 \$YMU 8403 Data delayed. In real time at http://is.gd/tGyI #futures #emini

TABLE 1. Four sample formats of Twitter messages (*Tweets*) used in this experiment. Note that the characters \$ and # are used as special markers to signify indexable content. In this case \$ leads in a ticker symbol and # a keyword (or *hashtag*). Note the use of an *URL Shortening Service* to compact the advertised hyperlink back to the Giller Investments website. This is common practice on Twitter due to the constraint of fitting the message into the 140 character hard limit.

t	n_t	s_t	$\hat{\mu}_t$
2009:04:20	2	1	6.43
2009:04:21	2	-1	7.03
2009:04:22	2	-1	7.82
2009:04:23	2	-1	8.60
2009:04:24	3	1	9.65
2009:04:27	4	-1	10.16
2009:04:28	5	-1	10.94
2009:04:29	9	1	12.06
2009:04:30	10	-1	12.50
2009:05:01	11	-1	13.29
2009:05:04	15	1	14.47
2009:05:05	15	1	15.27
2009:05:06	15	-1	15.63
2009:05:07	14	-1	16.41
2009:05:08	15	-1	17.19
2009:05:11	17	-1	17.97
2009:05:12	17	-1	18.76
2009:05:13	19	1	20.10
2009:05:14	18	-1	20.32
2009:05:15	19	-1	21.10
2009:05:18	21	1	22.51
2009:05:19	21	-1	22.66
2009:05:20	24	-1	23.45
2009:05:21	26	1	24.92
2009:05:22	27	-1	25.01
2009:05:25	29	-1	25.79
2009:05:26	30	1	27.33
2009:05:27	31	1	28.14
2009:05:28	32	-1	28.13
2009:05:29	30	-1	28.92
2009:06:01	37	1	30.55
2009:06:02	30	-1	30.48
2009:06:03	31	-1	31.26
2009:06:04	31	-1	32.04
2009:06:05	35	-1	32.82
2009:06:08	36	-1	33.61
2009:06:09	36	-1	34.39
2009:06:10	36	-1	35.17
2009:06:11	38	-1	35.95
2009:06:12	40	-1	36.73
2009:06:15	42	1	38.59
2009:06:16	44	-1	38.29
2009:06:17	46	-1	39.08
2009:06:18	46	-1	39.86
2009:06:19	49	-1	40.64

TABLE 2. Raw data used in the analysis of Section 4.