

MS&E 444 Final Project Report

Using Options Trading Data to Algorithmically Detect Insider Trading

Instructor: Prof. Kay Giesecke

TA: Benjamin Armbruster

Group Members:¹

Youdan Li

Elaine Ou

Florin Ratiu

Pawit Sangchant

Yantao Shen

June 14, 2007

¹ We highly appreciate the guidance and help by Prof. Kay Giesecke and TA Benjamin Armbruster. Your encouragement helped us overcome the difficulty in the project.

CONTENTS

1. Motivation and Introduction	3
2. Model	3
3. Data Analysis and Processing	4
4. Case Studies	6
5. Summary and Future Work	12
References	14

1. Motivation and Introduction

Options trading statistics are often used as indicators of current trends in an underlying stock. Individuals privileged with insider information can profit from their knowledge by discretely trading in stock options. There are a number of strategies that can be used in order to maximize profits and minimize detection. This can be exhibited in the options trading volume, put/call ratios, implied volatility, or other variables. In Mike Lipkin's 2006 talk, "Sherlock Trader, Takeover Sleuth", some strategies are described that would be particularly attractive to insiders who are trading stock options. These strategies can leave telltale signs in options trading volumes and implied volatilities, and, by identifying these signs, it can be possible to predict significant news events that are about to occur within a certain company.

In McMillan's book, the author points out that insiders wishing to exploit their information would be most likely to buy stock options because that results in the most highly leveraged gains. McMillan discusses indicators that would be present in stock option volume, option prices, implied volatility, and the put/call ratio.

Lipkin focuses on mergers and acquisitions as his target news events because those are definitive circumstances that will affect a stock's value in a predictable manner. He presents a number of cases where a subjective analysis of options volumes and implied volatility yields an obvious indication of insider trading.

In our work, we focus on trends present in stock option volume and implied volatility as indicators of impending news announcements. Because it is most useful to be able to predict a news event before it actually occurs, it would be necessary to analyze the options series of large numbers of companies. More importantly, the results of analysis have to be statistically testable to be meaningful. This would require immense amounts of data processing and a more rigorous framework than just simple ocular tests. Our contribution in this project is first to apply the binary choice model from Econometrics to account for the relationships between option trading and abnormal returns in stock prices and second to develop a semi-automated system to process options data in order to generate a predictor of a future news announcement. We describe a model based on a number of case studies, and how it can be used to automatically analyze and process options trading data for any given company.

2. Model

To investigate this problem, our model makes the following assumptions:

- (1) If there is any insider trading, the trading data must have some strong correlation with future abnormal returns in stock market;
- (2) Insider trading mainly happens within a relatively short period before the event

- announcement or day of abnormal returns;
- (3) The distribution function of error term is correct;
 - (4) Trading strategies are correctly built into model specification;
 - (5) Actively traded options are most relevant to future events.

We model the relationship between the option trading statistics and abnormal stock price returns using the Probit model.

$$y_t = \Phi(x_t' \beta) + u_t$$

where,

$\Phi(\cdot)$ is cumulative standard normal distribution;

y_t is a binary variable taking 1 or 0;

x_t is a vector of explanatory variables;

x_{t+1} is adapted to some filtration (F_t) ;

x_t is an integrated time series possibly of ARIMA type;

β : is the coefficients vector including a constant term ;

(u_t, F_t) is a Martingale Difference Sequence.

In this model, $y_i=0$ if no abnormal return, and $y_i=1$ if there is an abnormal return. x_i includes constants and lags in volume or implied volatilities of call/put options. The maximum likelihood method is used to find the coefficients (β), thus the asymptotic is

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} MN(0, V)$$

where $MN(0, V)$ is the mixed normal distribution. This means a class of distribution being normal conditional on a set of random variables. V is the relevant covariance matrix. Since Wald tests of restrictions on β have asymptotic chi-squared distributions, the statistical inference can proceed in the usual manner. Many econometric software packages can perform the Probit model estimation.

3. Data Analysis and Processing

While much of the previous work in this area focuses on closely examining “cherry-picked” examples of option trading that have been associated to insider information in finance folklore, we have tried to maximize the level of automation in the process of selecting the data samples. Automation and fast processing of large

option series are requirements for any system that tries to analyze data online. Our system's capabilities come relatively close to that and are described below. Note that we have used the OptionMetrics database for options data (available only from 01/1996 to 04/2006) and the CRSP database for historical stock price data.

Our first step was to look at historical data and detect all the particularly unusual stock price returns (more than 4 standard deviations away from the mean) and retrieve the relevant option series (up to 100 days) leading to the day of that very significant return. We used the (adjusted for splits) historical stock prices from CRSP and modeled the distribution of stock returns for each company as a normal distribution. Our motivation for this approach was the fact that insider trading is associated with very significant changes in the stock price and traces of someone who is trying to capitalize on this information might be found in the options series data that lead to the date of the significant return. For each active trading date, our system extracted the most traded 3 calls series and 3 puts series for each strike price. In each case, we were interested in the option's daily volume and implied volatility. To each day, we also associated the binary variable of the Probit model, and this was set to 1 when the daily return significantly deviated from the mean. This setup allowed us to quickly retrieve the "interesting" events in the stock price data as well as the most traded options around those significant events, to test our model on them.

This procedure can be easily adapted to make predictions on the stock return "tomorrow" – we just extract the most traded options series over the last 100 days leading all the way to the current date.

The next step in our automated process is to take the options series obtained by the procedure above and test it using the Matlab Probit package. This involves using different lags for each variable (strike price, expiration date)x(day, volume, implied volatility) series and picking the ones that have significant values. However, the automated program can only do preliminary analysis and selection due to the fact that different companies have different kinds of events and different trading strategies may be applied in each circumstance. In-depth analysis is still needed to find out more interesting and significant results.

We have also measured some statistics that relate to the hypothesis that unusual put/call volumes are indeed main indicators of significant future stock returns. For each daily trading volume more than a standard deviations from the average daily trading volume over the previous 100 days, we computed the sample probability that the stock return would also get to more than b standard deviations from its historical mean sometime over the next 15 days. Results for some companies are in the tables below:

MSFT		b – # stdev in stock	# ex
------	--	------------------------	------

		2	3	4	
<i>a</i> – # stdev in trading volume	2	0.355	0.108	0.03	166
	3	0.364	0.135	0.02	74
	4	0.354	0.129	0.06	31

AAPL		<i>b</i> – # stdev in stock			# ex
		2	3	4	
<i>a</i> – # stdev in trading volume	2	0.406	0.213	0.04	145
	3	0.380	0.226	0.03	84
	4	0.377	0.244	0.02	45

4. Case Studies

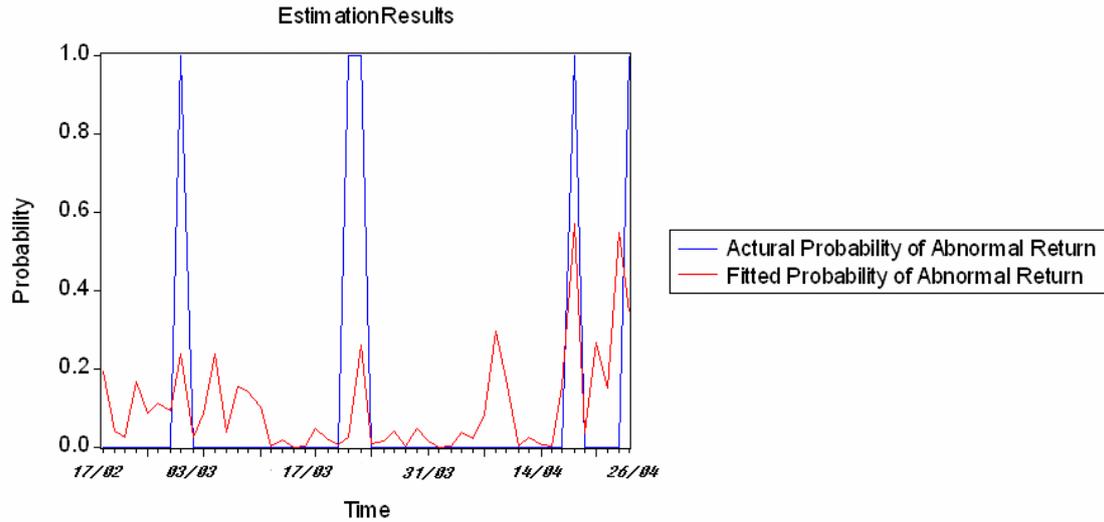
(a) Estimation results for FORE

Our analysis gives some support to Lipkin's conclusion by more rigorous methods with the same trading strategy.

Variable	Coefficient	Std. Error	z-Statistic	Prob.
Constant	43.33	27.04	1.60	0.11
IV11 with 30 days lag	-34.71	17.87	-1.94	0.05
IV12 with 30 days lag	23.13	15.07	1.53	0.12
IV22 with 20 days lag	-52.65	31.29	-1.68	0.09

The signs of coefficients of IVs are correct; The coefficients of IV11 and IV22 are statistically significant at 95% and 90% levels respectively.

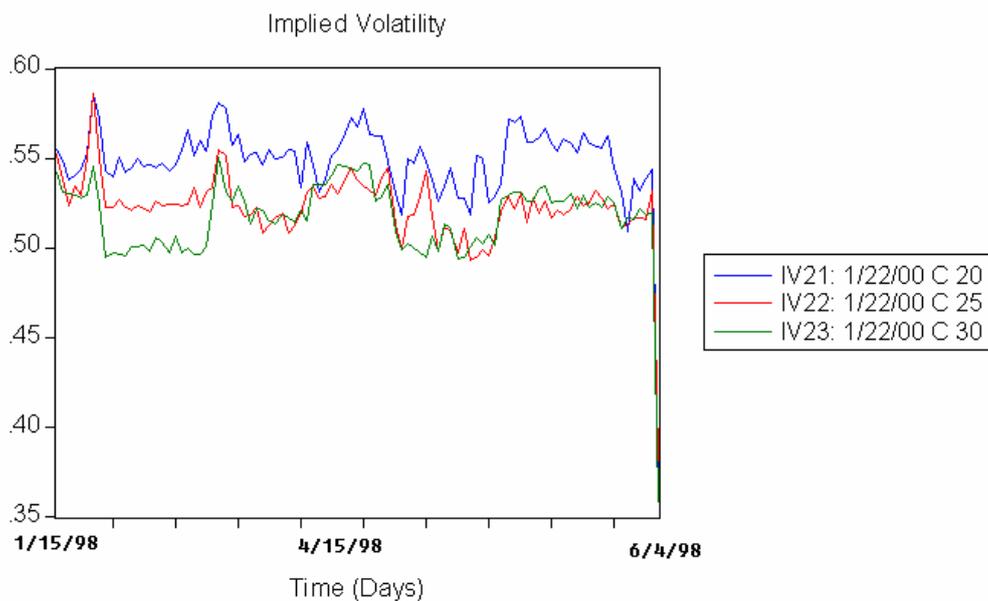
The following figure shows the fitted and actual probability of abnormal returns. Our model predicts the main peaks of actual probability.



(b) Estimation results for DIGI

1. On June 4, 1998, Alcatel announced its acquisition of DSC (DIGI);
2. Lipkin didn't find evidence about insider trading;
3. Our analysis is not completely consistent with Lipkin's conclusion since the implied volatilities seemed to be correlated with abnormal returns statistically.

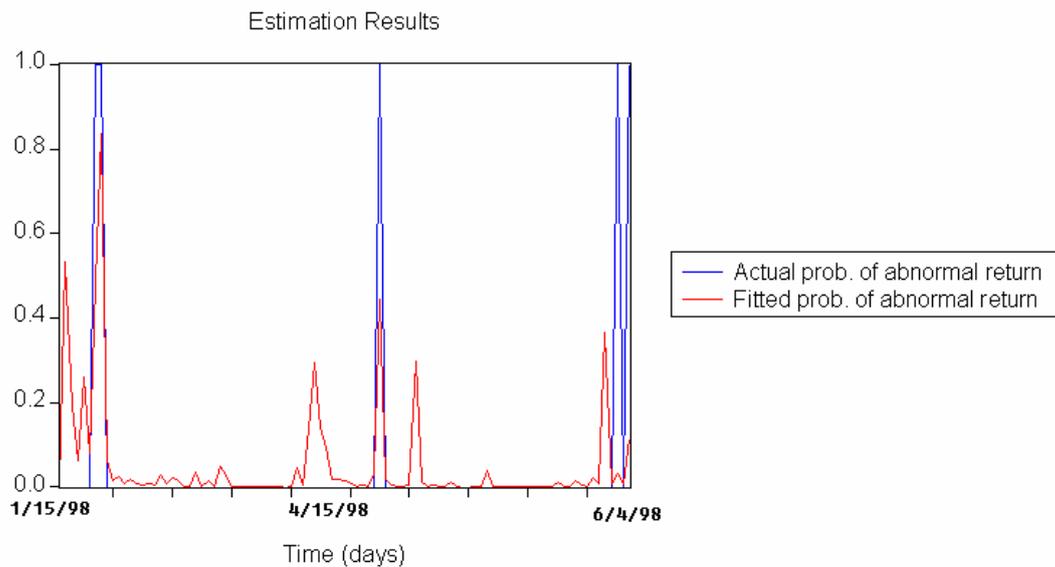
The following figure shows the implied volatility of our chosen options series. All three series fell dramatically on the event announcement day. Before that date, no obvious fluctuation could be observed for any of the three series.



In the following table, the estimated coefficients of IV21 and IV22 with one day lag showed the significant correlation with the major M&A event on 6/4/1998.

Variable	Coefficient	Std. Error	z-Statistic	Prob.
Constant	-14.02793	12.54665	-1.118062	0.2635
IV21 with one day lag	-78.41029	33.99238	-2.306702	0.0211
IV22 with one day lag	99.33289	35.60284	2.790027	0.0053
IV23 with one day lag	5.031176	23.15941	0.217241	0.828

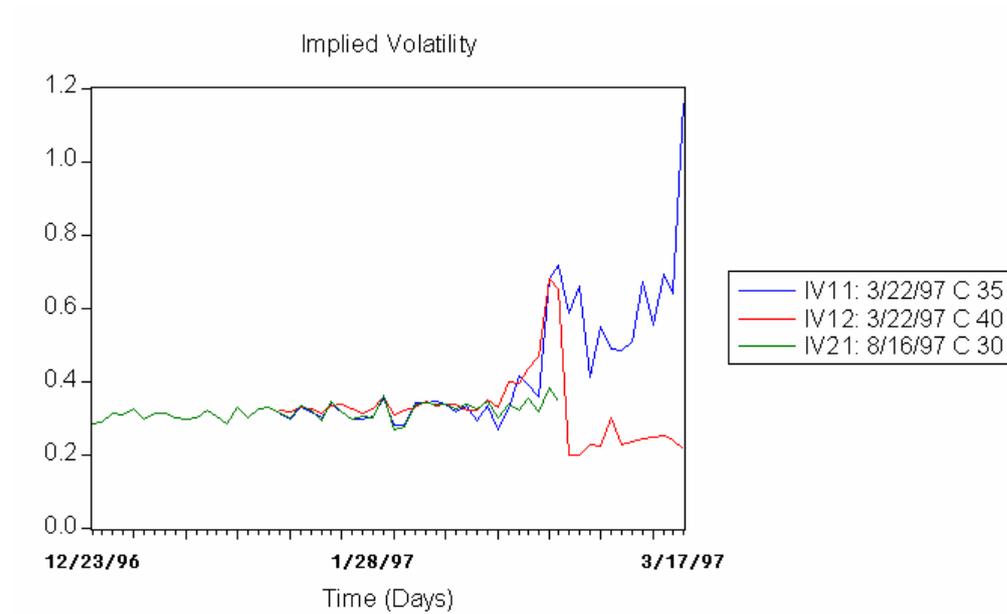
The figure below shows the quality of our model prediction. The model predicts two peaks correctly.



(c) Estimation results for COFD

1. On March 17, 1997, Summit Bancorp announced its acquisition of Collective Bank Corp (COFD);
2. Lipkin found evidence about two kinds of trading, sure speculation and insider trading;
3. Our analysis is partly consistent with Lipkin's conclusion.

The following graph shows the trends of three volatility series. Both IV11 and IV12 dropped before the event. On the event announcement day, IV11 rose dramatically, while IV12 didn't change much.

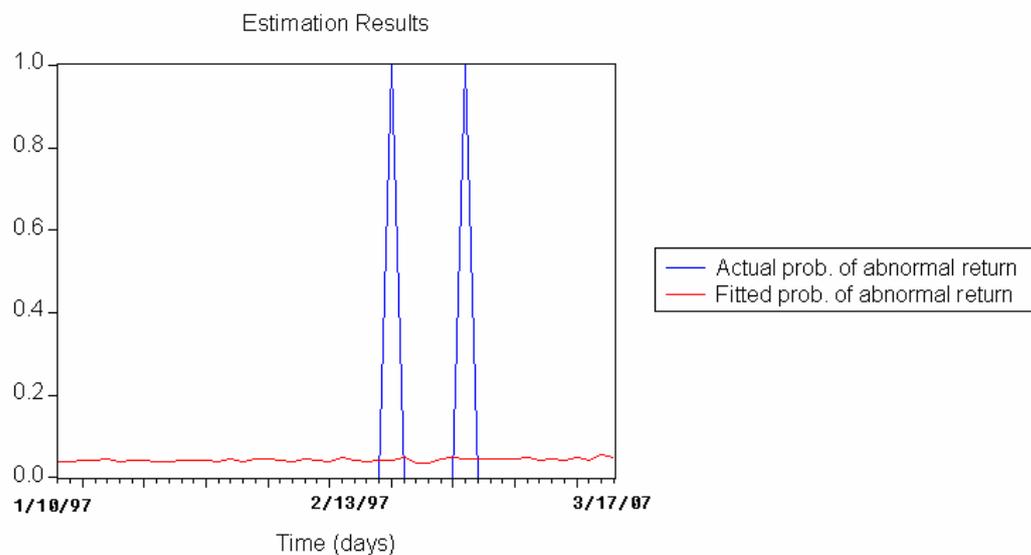


The results of our analysis are shown in the following two tables with different function specifications.

Variable	Coefficient	Std. Error	z-Statistic	Prob.
Constant	-2.28396	4.722167	-0.483668	0.6286
IV21with 12 days lag	1.78346	14.65696	0.12168	0.9032

No significant correlation was found between IV21 with 12-day-lag and the probability of the abnormal returns.

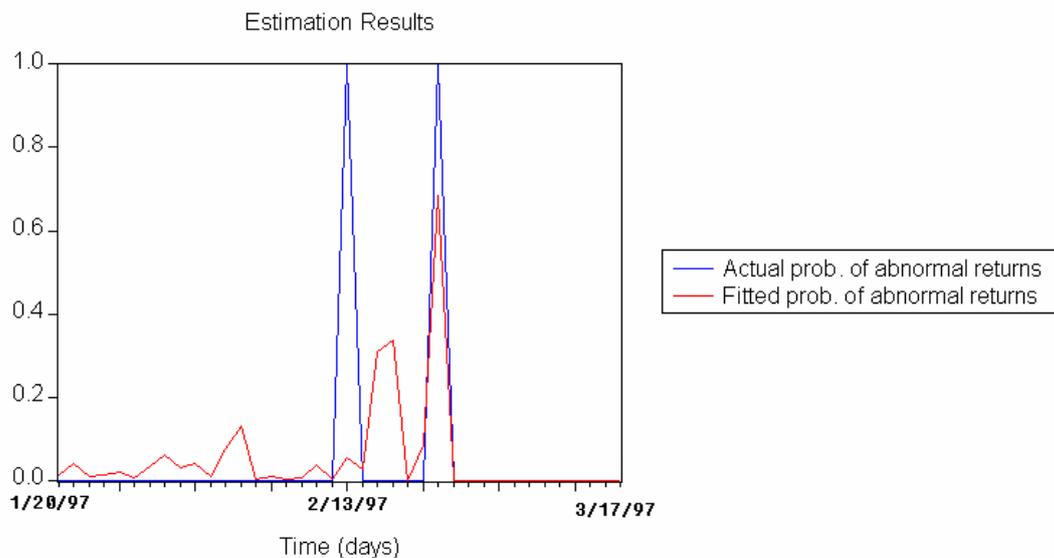
The estimation graph shows that IV21 with 12-day-lag doesn't explain the abnormal returns well.



Variable	Coefficient	Std. Error	z-Statistic	Prob.
Constant	-1.481679	4.051724	-0.365691	0.7146
IV11 with 2 days lag	-29.08965	21.3908	-1.359914	0.1739
IV12 with 2 days lag	26.42054	14.34046	1.842377	0.0654

Significant correlations were found between IV12 with 2-day-lag and the probability of the abnormal returns.

From the following graph, our model predicts one peak fairly well.



(d) Estimation results for MSFT

We tried to apply our model to a more general case, Microsoft. First, we chose the 100-day window period that had the greatest number of days with abnormal returns out of the total of 11 periods. (The abnormal return for this case is at least 3 standard deviations away from the historical mean.) It is the period from 9/12/2000-1/18/2001. We used this period to predict the abnormal return that occurred on 1/19/2001. After examining many options series, we discovered that by using the 1/18/2003 C 120, 1/19/2002 C 100, and 1/19/2002 C 75 series, we obtained the best predicted value for 1/19/2001. We then tried to adjust the lag number used in our model to find the best one given these three options series. The result was that a 2-day lag was the best for this case. After we obtained the best set of options series and the lag number, we tried to measure the sensitivity of the model. The results are summarized in the two tables below:

Lag	Predicted Value
-----	-----------------

1	0.1070
2	0.9773
3	0
4	0.1118
5	0

Table MSFT1: Same Series (1, 4, 6) Different Lag

Series	Predicted Value
1,4,5	0.0023
1,4,6	0.9973
1,5,6	0.9371
2,4,6	0.1128
3,4,6	0.1136

Table MSFT2: Different Series Same Time Lag(2)

Note: The series numbers are defined as follows:

1: 1/18/2003 C 120 2: 1/18/2003 C 125 3: 1/18/2003 C 100

4: 1/19/2002 C 100 5: 1/19/2002 C 70 6: 1/19/2002 C 75

Although we obtained a very good predicted value from Series 1, 4 and 6 with 2-day lags, the coefficients of the regression are insignificant. This can mean two things: that there is a trade-off between a good model and a good prediction value, or that the events that we consider do not really have any statistical significance. The summary of the coefficients and their corresponding statistical measures are shown below:

Coefficient	T-stat	Prob(P > t)
-------------	--------	---------------

-11.1026	-0.0000	1.000
0.0015	0.0000	1.000
0.0002	0.0000	1.000
-0.1988	-0.0000	1.000

Table MSFT3: Results from Series 1, 4, 6 with lag number = 2

5. Summary and Future Work

The options market is very important to the early detection of insider trading since options are such thinly traded securities. This feature means that a small amount of insider trading has a large impact on the overall trading data. In this way, options trading data is quite sensitive to systematic insider trading behavior, compared with heavily traded stock data where the noise of normal trading easily suppresses the signals of insider trading. Two of the most important indicators of options trading are volume and implied volatility.

In previous research done by Lipkin (2006), the signatures of insider trading are identified by manually plotting data and ocular tests. This is a straightforward methodology and easy to understand. However, this methodology relies on subjective judgment and can't provide any statistically testable results. Also, it lacks the power to detect relatively minor signals existing in trading data. Our methodology adopts the Probit model from econometrics to investigate the statistical relationship between option trading and future major events (or abnormal returns in stock prices) systematically and more rigorously. Finally, we develop a semi-automated system to process data and conduct analysis.

Our methodology has apparent advantages. First, it has great flexibility. Different trading strategies can be easily built into the model by adjusting explanatory variables and/or the combination of lags. Also it is not hard to change the assumption about the distribution of the error term to take into account complexity of reality. The analysis window is very easy to adjust in our system. Second, the reliability is relatively high if the data set is large enough since the estimators in our model are consistent asymptotically. Therefore, the results of analysis are statistically testable. Moreover, the model can be used with nonstationary time series data. In the setting of nonstationarity, the convergence rate is $n^{1/4}$ in contrast with $n^{1/2}$ in stationary cases.

Third, it has high efficiency. It can automatically process raw data and conduct preliminary analysis conveniently.

However, when using the system, we still need to pay attention to the following points. In-depth analysis for each company is usually required given the complexity of strategies and companies. It is possible to find a very well-fitting model but the estimation is very “sensitive” to model specification, which means using the same set of series but different lags can yield very different results, or using the same lag but one series different from the initial set can also yield different results. Therefore, we need trials and errors. Also, to tell whether insider trading happened, we need to exhaust as many strategies as possible.

Future work can include two aspects. One is to improve the degree of automation of the system. The other is to use more sophisticated econometrics model in our analysis. For instance, a more complicated limited dependent variable model can be used to tell if abnormal returns are positive or negative.

Reference

- [1] Lipkin, Mike. Dec. 1, 2006. "Sherlock Trader: Take-Over Sleuth".
- [2] McMillian, L., McMillan on Options, John Wiley & Sons, 1996.
- [3] Donoho, Steve. 2004. "Early Detection of Insider Trading in Option Markets".