Using Options Trading Data to Algorithmically Detect Insider Trading

Youdan Li
Elaine Ou
Florin Ratiu
Pawit Sangchant
Yantao Shen
Goal

- Predict jumps in stock price ("events")
- Using evidence of insider trading in options data
- Automated system
On April 26, 1999, GEC announced its acquisition of FORE System Inc (FORE);

Lipkin found some evidence about insider trading using this graph given a specific trading strategy;
Lipkin: Insider Strategies

- Sell long-term premium via calendars
  - e.g. +100Jun(35)C-100Nov(35)C, 32.5 is the stock price
  - Highly risky and suggests insider trading
  - Calendar spreads will be crushed

- Do the near-term 1-by-many for a credit
  - e.g. -50Jun(32.5)C+200Jun(35)C, 32.5 is the stock price
  - Very safe and suggests pure speculation
  - At-the-money’s implied volatility (IV) will be reduced; Next higher strike’s IV will be elevated
Overview

- Model
- Data analysis and processing
- Microsoft sample
- Summary and future work
- Q&A
Model Assumptions

- If there is any insider trading, the trading data must have some strong correlation with future abnormal returns in stock market;
- Inside trading mainly happened within a relatively short period before the event announcement or abnormal returns;
- Distribution function of error term is correct;
- Trading strategies are correctly built into model specification;
- Actively traded options are more relevant to future events.
Nonstationary Probit Model

\[ y_t = \Phi(x_t' \beta) + u_t \]

- \(\Phi(\cdot)\) 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;
- \(\beta\): is the coefficients vector including a constant term;
- \((u_t, F_t)\) is a Martingal Difference Sequence.
Model

- $Y_i=0$ if no abnormal return;
- $Y_i=1$ if abnormal return;
- $X_i$ includes constant and lags of volume and/or implied volatilities of call/put options;
- Maximum likelihood method is used to find coefficients (Beta), the asymptotic is

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

$MN$ is mixed normal distribution (being normal conditionally on a random variable);

$V$ is the relevant variance matrix.
Data Analysis

- use historical prices to create normal model of daily stock return
- For a given company, some data preprocessing is done automatically:
  - An abnormal return is defined as $1.96\sigma$ from the mean
  - For each day with a very significant return ($> 4$ stdev from mean)
    - Consider all options traded up to 100 days before the day
    - Extract the top 3 most traded put/call options for each possible expiration date
    - Output: trading volume, implied volatility, binary variable
Our estimation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>z-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>43.33</td>
<td>27.04</td>
<td>1.60</td>
<td>0.11</td>
</tr>
<tr>
<td>IV11 with 30 days lag</td>
<td>-34.71</td>
<td>17.87</td>
<td>-1.94</td>
<td>0.05</td>
</tr>
<tr>
<td>IV12 with 30 days lag</td>
<td>23.13</td>
<td>15.07</td>
<td>1.53</td>
<td>0.12</td>
</tr>
<tr>
<td>IV22 with 20 days lag</td>
<td>-52.65</td>
<td>31.29</td>
<td>-1.68</td>
<td>0.09</td>
</tr>
</tbody>
</table>

- Our analysis gave some support to Lipkin’s conclusion by more rigorous methods with the same trading strategy.

  - The signs of coefficients of IVs are correct;
  - The coefficients of IV11 and IV22 are statistically significant at 95% and 90% levels respectively.
Our estimation
An Example from Microsoft Case

- Define the abnormal return as $3 \sigma$ away from the historical mean.
- One abnormal return happened on 1/19/2001.
- Use three series of options’ volume traded in 100-day window period before that day for the prediction.
Series Used in the Analysis

- Series 1: 1/18/2003 C 120
- Series 2: 1/18/2003 C 125
- Series 3: 1/18/2003 C 100
- Series 4: 1/19/2002 C 100
- Series 5: 1/19/2002 C 70
- Series 6: 1/19/2002 C 75
## Results (Series: 1, 4, 6 & Time Lag = 2)

| 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     |
Same Series (1,4,6) with Different Lags

<table>
<thead>
<tr>
<th>Lag</th>
<th>Predicted Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1070</td>
</tr>
<tr>
<td>2</td>
<td>0.9773</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.1118</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>
Predicted Value with Different Sets of Series but with the Same Time Lag(2)

<table>
<thead>
<tr>
<th>Series</th>
<th>Predicted Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,4,5</td>
<td>0.0023</td>
</tr>
<tr>
<td>1,4,6</td>
<td>0.9973</td>
</tr>
<tr>
<td>1,5,6</td>
<td>0.9371</td>
</tr>
<tr>
<td>2,4,6</td>
<td>0.1128</td>
</tr>
<tr>
<td>3,4,6</td>
<td>0.1136</td>
</tr>
</tbody>
</table>
Summary

- **Great flexibility**
  - Trading strategies can be easily built in
  - Distribution function can be adjusted
  - Analysis window is flexible

- **Reliability**
  - Consistent estimator
  - Statistically testable results
  - Robust to nonstationarity of time series data

- **High efficiency**
  - Automatically process raw data
  - Convenient preliminary analysis
Summary

- In-depth analysis for each company is required
  - Possible to find a very well-fitting model;
  - The estimation is very “sensitive” to model specification;
  - 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;
  - Need trials and errors!

- Future work
  - Improve the degree of automation of the system