MS&E 444 Project 5

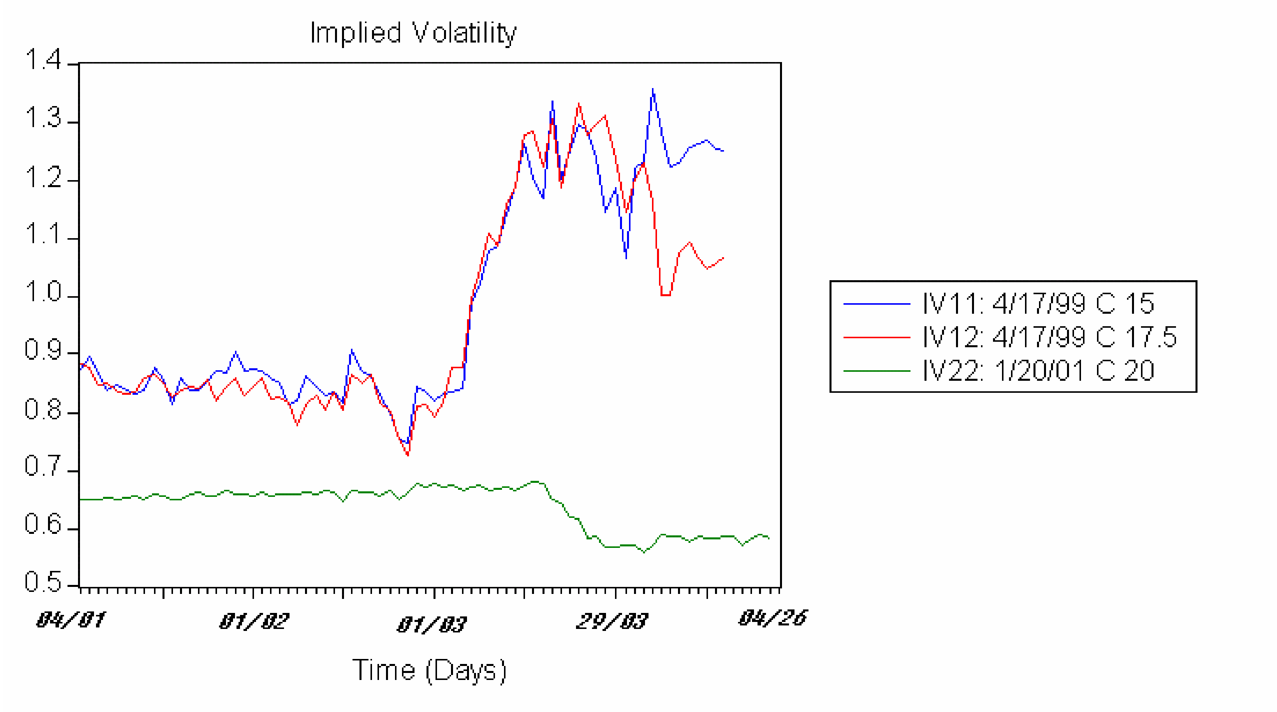
Using Options Trading Data to Algorithmically Detect Insider Trading

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

Lipkin Analysis: FORE



- 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;

β: 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

$$\sqrt[4]{n}(\hat{\beta}_n - \beta) \xrightarrow{d} MN(0,V),$$

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

V is the relevant variance matrix.

Data Analysis

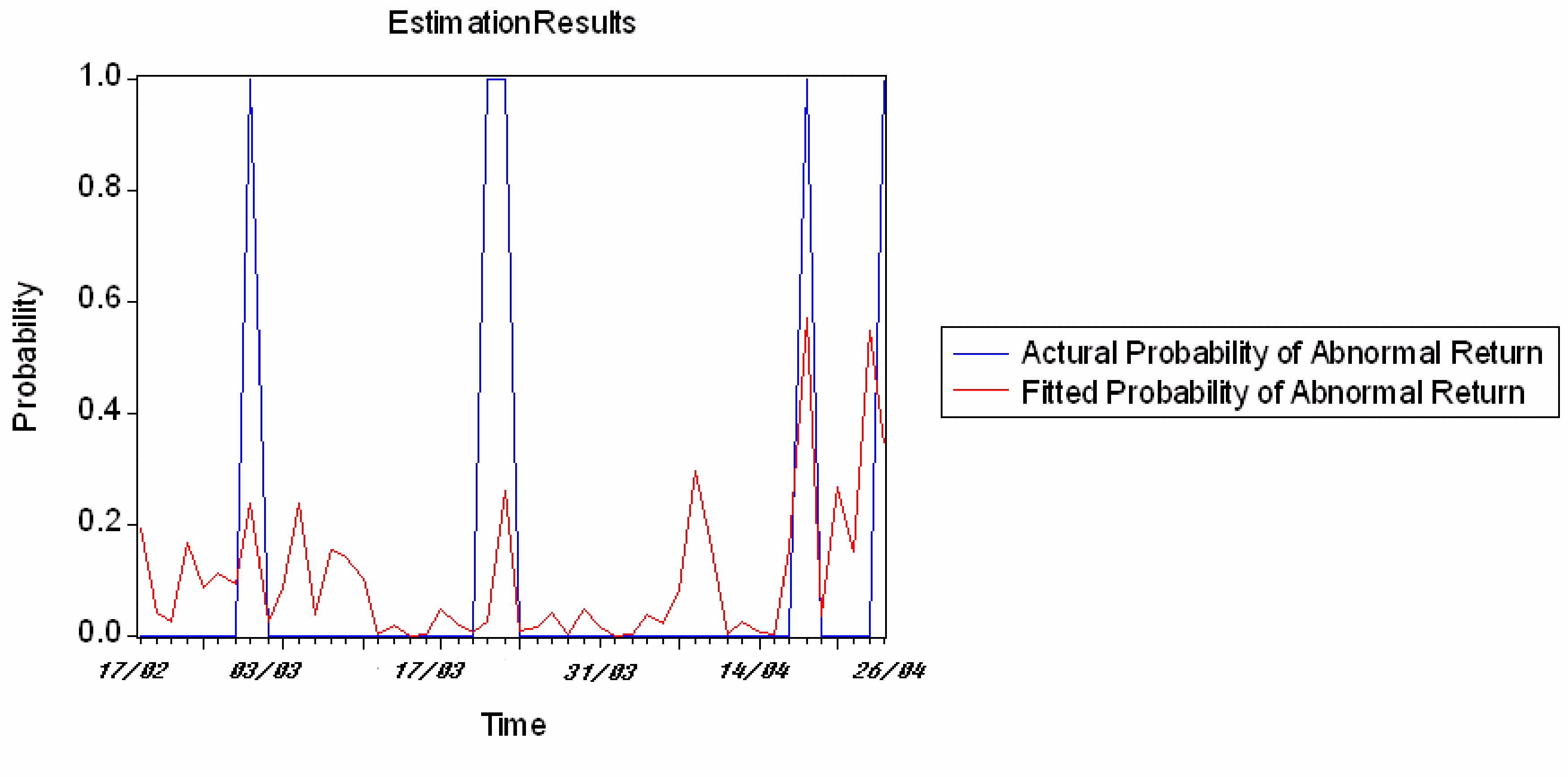
- use historical prices to create normal model of daily stock return
- for a given company, some data preprocessing is done automatically:
- An abnormal return is defined as 1.96 σ from the mean
- for each day with a very significant return (> 4 stdev from mean)
  - consider all options traded up to 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

Variable	Coefficient	Std. Error	z-Statistic	Prob.
Constant	43.33	27.04	1.60	0.11
IV11with 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

- 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 σ 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

Lag	Predicted Value
1	0.1070
2	0.9773
3	0
4	0.1118
5	0

Predicted Value with Different Sets of Series but with the Same Time Lag(2)

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

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

Q&A