LECTURE 8

Analysis of High-Frequency Financial Data & Market Microstructure

Market microstructure: Why is it important?

1. Important in market design & operation, e.g. to compare different markets (NYSE vs NASDAQ)
2. To study price discovery, liquidity, volatility, etc.
3. To understand costs of trading
4. Important in learning the consequences of institutional arrangements on observed processes, e.g.
   - Nonsynchronous trading
   - Bid-ask bounce
   - Impact of changes in tick size, after-hour trading, etc.
   - Impact of daily price limits (many foreign markets)

Nonsynchronous trading:
Key implication: may induce serial correlations even when the underlying returns are iid.
Setup: log returns $\{r_t\}$ are iid (μ, σ^2)
For each time index t, P(no trade) $= \pi$.
Cannot observe r_t if there is no trade.
What is the observed log return series r_t^o?
It turns out r_t^o is given in Eq. (5.1),

$$r_t^0 = \begin{cases} 0 & \text{with prob. } \pi \\ r_t & \text{with prob. } (1-\pi)^2 \\ r_t + r_{t-1} & \text{with prob. } (1-\pi)^2\pi \\ \vdots & \vdots \\ \Sigma_{i=0}^{k} r_{t-i} & \text{with prob. } (1-\pi)^2\pi^k \\ \vdots & \vdots \end{cases}$$

geometric sums and some tricks

One can use this relation to show that

$$\begin{aligned} \text{Var}(r_t^o) &= \sigma^2 + \frac{2\pi\mu^2}{1-\pi} \\ \text{Cov}(r_t^o, r_{t-j}^o) &= -\mu^2\pi^j, \quad j \geq 1. \end{aligned}$$

stretch of unobserved returns

r_t^o = cumulative value of r_t with previous

k	no trade	yes trade	r_t^o	prob
<0	t	na	0	π
0	na	$t, t-1$	r_t	$(1-\pi)^2$
1	$t-1$	$t, t-2$	$r_t + r_{t-1}$	$(1-\pi)^2\pi$
2	$t-1, t-2$	$t, t-3$	$r_t + r_{t-1} + r_{t-2}$	$(1-\pi)^2\pi^2$

etc.

Bid-ask bounce

Bid and ask quotes introduce **negative** lag-1 serial correlation.

Setup: simplest case of Roll(1984)

True price $P_t^* = \frac{P_a+P_b}{2}$ is unchanged over time, i.e. $P_t^* = P_{t-1}^*$

Story: market makers facilitate trades in some markets; make money by selling at higher price (P_a) than they buy (P_b); this creates the so called bid-ask bounce

But: the obs price P_t is subject to "friction" as follows:

$S = P_a - P_b$ is the bid-ask spread

$$P_t = P_t^* + \begin{cases} S/2 & \text{with prob. } 0.5 \\ -S/2 & \text{with prob. } 0.5 \end{cases}$$

Then, $P_t = P_t^* + \frac{S}{2} I_t$ and

$$\Delta P_t \equiv P_t - P_{t-1} = (I_t - I_{t-1})\frac{S}{2}$$

where I_t and I_{t-1} are independent binary variables with $P(I_i = 1) = 0.5$ and $P(I_i = -1) = 0.5$.

Note: $E(I_t) = 0$ and $\text{Var}(I_t) = 1$ for all t.

One can show that $\mathbb{E}(\Delta P_t) = 0$

$$\begin{aligned} \text{Var}(\Delta P_t) &= S^2/2 \\ \text{Cov}(\Delta P_t, \Delta P_{t-1}) &= -S^2/4 \\ \text{Cov}(\Delta P_t, \Delta P_{t-j}) &= 0, \quad j > 1. \end{aligned}$$

The result continues to hold if P_t^* follows a random walk model. That is, $P_t^* = P_{t-1}^* + e_t$ with $e_t \sim iid(0, \sigma_2^2)$.

High-Frequency Financial Data

Observations taken with time intervals 24 hours or less

Some examples:

1. Transaction (or tick-by-tick) data
2. 5-minute returns in FX
3. 1-minute returns on index futures and cash market

Some Basic Features of the Data:

1. Irregular time intervals
2. Leptokurtic or Heavy tails
3. Discrete values, e.g. price in multiples of tick size
4. Large sample size
5. Multi-dimensional variables, e.g. price, volume, quotes, etc.
6. Diurnal Pattern

An illustration: Consider the transaction-by-transaction data of Johnson and Johnson from October 4 to October 15, 2010. There are 418,855 intraday price changes. Original data are from NYSE TAQ.

Time plot and histogram of intraday price changes in consecutive trades: See Figure 2. The histogram indicates most transactions are without price change.

The number of transactions in 5-min time intervals: (a) Time plot and (b) ACF: See Figure 3. The ACF shows a clear diurnal pattern in trading intensity.

R demonstration

```
> da=read.table("taq-jnj-t-oct4t152010.txt",header=T)
> head(da)
      date hour minute second price volume
1 20101004    6     25     15 61.75    100
2 20101004    8     33     19 61.56    100
3 20101004    8     41      9 61.56    100
4 20101004    8     48     50 61.60    100
5 20101004    8     48     55 61.60    100
6 20101004    8     49      4 61.60    100
> source("hfchg.R")   ### R script to compute price change
> m1=hfchg(da)
number of trading days:  10
> names(m1)
[1] "pchange"  "duration" "size"
> par(mfcol=c(2,1)); idx=c(410000:418854)
> plot(m1$pchange,type='l',ylab='change') #plot(idx,m1$pchange[idx],type='l',ylab='pch')
> hist(m1$pchange, nclass=400, xlim=c(-0.04,0.04)) ### May use xlim=c(-0.06,0.06)

> source("hfntra.R")  # R script to tabulate number of transactions in a given
                                time interval (measured in minutes).
> m1=hfntra(da,5)
> names(m1)
[1] "ntrad"
```

Frequencies of price change

Cents	≤ -2	$[-2,-1)$	$[-1,0)$	0	(0,1]	(1,2]	≥ 2
Counts	915	6768	49976	304066	49552	6655	922
Percentage	0.218	1.616	11.932	72.595	11.830	1.589	0.220

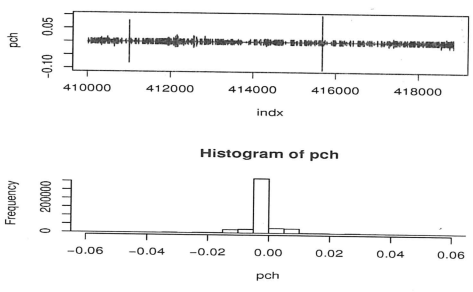

Figure 2: Time plot and histogram of intraday price changes in consecutive trades for JNJ stock from October 4 to October 15, 2010. Only a small portion of the price changes (418854 data points) is shown in the upper plot.

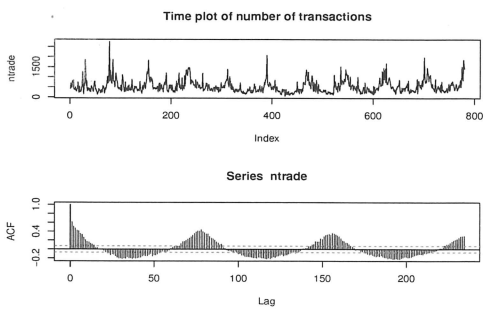

Figure 3: Time plot of the number of transactions in 5-min time intervals and its sample ACF for JNJ stock from October 4 to October 15, 2010.

Econometric models used in the literature

1. Duration models, e.g. autoregressive conditional duration (ACD) models.
2. Models for price changes (ADS Model)
3. Models for bid and ask quotes

We focus on simple models for price change. (ADS) first. Later will do ACD.

Price Change: Discrete values

- Ordered probit model: Hauseman, Lo, & MacKinlay (1992)
- ADS model: Rydberg & Shephard (1998), McCulloch & Tsay (2000)

1 ADS Decomposition Models

A simple ADS decomposition:

- Price $P_t = P_0 + \Sigma_i^{N(t)} C_i$
- Number of transactions in [0,t]: $N(t)$
- $C_i = A_i D_i S_i$ $= P_{t_i} - P_{t_{i-1}}$ = price change

where: $\begin{cases} t_i = \text{time of } i^{th} \text{ transaction} \\ P_{t_i} = \text{price @ } t_i \end{cases}$

  - Action:
$$A_i = \begin{cases} 1 & \text{if } C_i \neq 0 \\ 0 & \text{otherwise} \end{cases}$$
  - Direction, given $A_i = 1$:
$$D_i = \begin{cases} 1 & \text{if } C_i > 0 \\ -1 & \text{if } C_i < 0 \end{cases}$$
  - Size, given $A_i = 1$ and D_i: multiple of tick size . Size = S_i = positive integer

- Can be estimated by the logistic regression method

  logit : $\ln[p/(1-p)]$ = linear function of explanatory variables $\longrightarrow$ $p = P(A=1)$.

A brief introduction of logistic regression: A case of two explanatory variables X and Z. The probability p_i is related to the observed values $X = x_i$ and $Z = z_i$ via the equation

$$\ln[p_i/(1-p_i)] = \beta_0 + \beta_1 x_i + \beta_2 z_i. \quad = \text{logit}(p) = l$$

This is equivalent to

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 z_i)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 z_i)}. \quad = \frac{e^l}{1+e^l}$$

It has many applications, e.g. probability of approving a loan based on the social and economic variables of an applicant.
We can use the command `glm` in R to perform logistic regression analysis.

Model specification of ADS models:

- Action A_i: Governed by a logistic regression

$$P(A_i = 1|F_{i-1}) = \text{logit}(F_{i-1})$$

- Direction given $A_i = 1$:

$$P(D_i = 1|F_{i-1}, A_i = 1) = \text{logit}(A_i, F_{i-1})$$

- Size given $A_i = 1$ and D_i:

$$P(S_i = s|A_i = 1, D_i = 1, F_{i-1}) \sim 1 + g(\lambda_{u,i})$$
$$P(S_i = s|A_i = 1, D_i = -1, F_{i-1}) \sim 1 + g(\lambda_{d,i})$$

where $g(.)$ denotes a Geometric distribution and $\lambda_{j,i}$ is governed by a logistic equation:

$$\ln(\frac{\lambda_{j,i}}{1-\lambda_{j,i}}) = \text{linear function of } F_{i-1}, A_i = 1, D_i.$$

Likelihood function:

$$P(C_i = s|F_{i-1}) =$$
$$P(S_i = s|A_i = 1, D_i, F_{i-1})P(D_i|A_i = 1, F_{i-1})P(A_i = 1|F_{i-1}).$$

An example: IBM data 59,775 observations. (Example 5.2 of the textbook.)

- Predictors: $\{A_{i-1}, D_{i-1}, S_{i-1}, V_{i-1}, x_{i-1}, BA_i\}$ (✓ ✓ ✓ × × ×)
  1. V_{i-1}: volume of the previous trade (divided by 1000)
  2. x_{i-1}: previous duration
  3. BA_i: the prevailing bid-ask spread

  } were not sig. in model

- Model:
  1. Action: $P(A_i|F_{i-1}) = p_i$, $\text{logit}(p_i) = \beta_0 + \beta_1 A_{i-1}$
  2. Direction: $P(D_i = 1|A_i = 1, F_{i-1}) = \gamma_i$, $\text{logit}(\gamma_i) = \delta_0 + \delta_1 D_{i-1}$

$$\Rightarrow P(D_i=1 \mid A_i=1, D_{i-1}) = \frac{e^{\delta_0 + \delta_1 D_{i-1}}}{1 + e^{\text{same}}}$$

  3. Size: $\text{logit}(\lambda_{j,i}) = \theta_{j,0} + \theta_{j,1} S_{i-1}$ with $j = d$ or u.

$$\Rightarrow P(S_i=s \mid A_i=1, S_{i-1}, D_i=j) \sim 1 + g(\lambda_{j,i})$$

where

$$\log\left(\frac{\lambda_{j,i}}{1-\lambda_{j,i}}\right) = \theta_{j,0} + \theta_{j,1} S_{i-1}$$

- Results:

Parameter	β_0	β_1	δ_0	δ_1
Estimate	−1.057	0.962	−0.067	−2.307
Std.Err.	0.104	0.044	0.023	0.056
Parameter	$\theta_{u,0}$	$\theta_{u,1}$	$\theta_{d,0}$	$\theta_{d,1}$
Estimate	2.235	−0.670	2.085	−0.509
Std.Err.	0.029	0.050	0.187	0.139

(See Table 5.6 of text.)

Implication

1. Prob of price change:

$$p_0 \equiv \quad P(A_i = 1|A_{i-1} = 0) = 0.258 = \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

$$p_1 \equiv \quad P(A_i = 1|A_{i-1} = 1) = 0.476. = \frac{e^{\beta_0+\beta_1}}{1+e^{\text{same}}}$$

2. Interpretation: **Odds ratio**

Because A_{i-1} is also a binary variable, we have a 2×2 table:

Outcome A_i	Independent variable A_{i-1}	
	$A_{i-1} = 1$	$A_{i-1} = 0$
$A_i = 1$	$P(A_i = 1) = \frac{\exp[\beta_0+\beta_1]}{1+\exp[\beta_0+\beta_1]}$	$P(A_i = 1) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$
$A_i = 0$	$P(A_i = 0) = \frac{1}{1+\exp[\beta_0+\beta_1]}$	$P(A_i = 0) = \frac{1}{1+\exp(\beta_0)}$.

simplify notation →

	$A_{i-1}=1$	$A_{i-1}=0$
$A_i=1$	p_1	p_0
$A_i=0$	$1-p_1$	$1-p_0$

e.g., $p_0 = \frac{e^{\beta_0}}{1+e^{\beta_0}}$

Odds Ratio: Row one divided by Row 2, then Column 1 divided by Column 2.

$$OR = e^{\beta_1}, \quad \text{or} \quad \beta_1 = \ln(OR).$$

Proof:

$$OR_0 \equiv \frac{p_0}{1-p_0} = e^{\beta_0}$$

$$OR_1 \equiv \frac{p_1}{1-p_1} = e^{\beta_0+\beta_1}$$

$$\Rightarrow OR = \frac{OR_1}{OR_0} = e^{\beta_1}$$

Thus: odds of $A_i=1|A_{i-1}=1$ are e^{β_1} times odds of $A_i=1|A_{i-1}=0$.

3. Direction of price change:

$$P(D_i = 1|F_{i-1}, A_i) = \begin{cases} 0.483 & \text{if } D_{i-1} = 0, \text{ i.e. } A_{i-1} = 0 \\ 0.085 & \text{if } D_{i-1} = 1,\ A_i = 1 \\ 0.904 & \text{if } D_{i-1} = -1,\ A_i = 1 \end{cases}$$

Bid-ask bounce

4. Weak evidence of price change cluster: price increases

$$S_i|(D_i = 1) \sim 1 + g(\lambda_{u,i}), \quad \lambda_{u,i} = 2.235 - 0.670 S_{i-1}.$$

↑ logit(·)

R demonstration: **glm** stands for generalized linear model.

```
> da=read.table("ibm91-ads.txt",header=T)
> da1=read.table("ibm91-adsx.txt",header=T)
> head(da)
  Ai Di Si
1  0  0  0
2  0  0  0
3  0  0  0
4  0  0  0
5  1  1  1
6  1 -1  1
> head(da1)
  Vim1 Durm1   BAi Aim1 Dim1 Sim1
1    8   0.4 0.125    0    0    0
2    0   0.1 0.370    0    0    0
3    1   1.0 0.125    0    0    0
4    5   0.1 0.125    0    0    0
5    4   0.1 0.625    0    0    0
6   62   1.0 0.625    1    1    1
> Ai=da$Ai
> Di=da$Di
> Aim1=da1$Aim1
> Dim1=da1$Dim1
> m1=glm(Ai~Aim1,family=binomial)  % fit a linear logistic regression
> summary(m1)
```

$\beta_0 \rightarrow$

$\beta_1 \rightarrow$

```
Call:
glm(formula = Ai ~ Aim1, family = binomial)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.05667    0.01142  -92.55   <2e-16 ***
Aim1         0.96164    0.01827   52.62   <2e-16 ***
---

> di=Di[Ai==1]
> dim1=Dim1[Ai==1]
> di=(di+abs(di))/2  % Transform di into a binary variable
> m2=glm(di~dim1,family=binomial)
> summary(m2)
Call:
glm(formula = di ~ dim1, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.06663    0.01728  -3.855 0.000116 ***
dim1        -2.30693    0.03595 -64.171  < 2e-16 ***
---
    Null deviance: 27335  on 19717  degrees of freedom
Residual deviance: 20039  on 19716  degrees of freedom
```

logit(S_i)

S_0

S_1

logit(p_i)

2 Ordered Probit Model

Let y_i^* be the unobservable price change of the asset under study (i.e., $y_i^* = P_{t_i}^* - P_{t_{i-1}}^*$), where P_t^* is the *virtual* price of the asset at time t. The ordered probit model assumes that y_i^* is a continuous random variable and follows the model

$$y_i^* = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \tag{1}$$

where $\boldsymbol{x}_i$ is a p-dimensional row vector of explanatory variables available at time t_{i-1}, $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector, $E(\epsilon_i|\boldsymbol{x}_i) = 0$, $\mathrm{Var}(\epsilon_i|\boldsymbol{x}_i) = \sigma_i^2$, and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. The conditional variance σ_i^2 is assumed to be a positive function of the explanatory variable $\boldsymbol{w}_i$ – that is,

$$\sigma_i^2 = g(\boldsymbol{w}_i), \tag{2}$$

where $g(.)$ is a positive function. For financial transactions data, $\boldsymbol{w}_i$ may contain the time interval $t_i - t_{i-1}$ and some conditional heteroscedastic variables. Typically, one also assumes that the conditional distribution of ϵ_i given $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ is Gaussian.

Suppose that the observed price change y_i may assume k possible values. In theory, k can be infinity, but countable. In practice, k is finite and may involve combining several categories into a single value. For example, we have $k = 7$ in Table 1, where the first value "< -2 cents" means that the price drops more than 2 cents. We denote the k possible values as $\{s_1, \ldots, s_k\}$. The ordered probit model postulates the relationship between y_i and y_i^* as

$$y_i = s_j \quad \text{if} \quad \alpha_{j-1} < y_i^* \leq \alpha_j, \quad j = 1, \cdots, k, \tag{3}$$

where α_j are real numbers satisfying $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{k-1} < \alpha_k = \infty$. Under the assumption of conditional Gaussian distribution, we have

$$\begin{aligned} P(y_i = s_j|\boldsymbol{x}_i, \boldsymbol{w}_i) &= P(\alpha_{j-1} < \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i \leq \alpha_j|\boldsymbol{x}_i, \boldsymbol{w}_i) \\ &= \begin{cases} P(\boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i \leq \alpha_1|\boldsymbol{x}_i, \boldsymbol{w}_i) & \text{if } j = 1, \\ P(\alpha_{j-1} < \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i \leq \alpha_j|\boldsymbol{x}_i, \boldsymbol{w}_i) & \text{if } j = 2, \ldots, k-1, \\ P(\alpha_{k-1} < \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i|\boldsymbol{x}_i, \boldsymbol{w}_i) & \text{if } j = k, \end{cases} \\ &= \begin{cases} \Phi\left[\frac{\alpha_1 - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i(\boldsymbol{w}_i)}\right] & \text{if } j = 1, \\ \Phi\left[\frac{\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i(\boldsymbol{w}_i)}\right] - \Phi\left[\frac{\alpha_{j-1} - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i(\boldsymbol{w}_i)}\right] & \text{if } j = 2, \ldots, k-1, \\ 1 - \Phi\left[\frac{\alpha_{k-1} - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i(\boldsymbol{w}_i)}\right] & \text{if } j = k, \end{cases} \end{aligned} \tag{4}$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal random variable evaluated at x, and we write $\sigma_i(\boldsymbol{w}_i)$ to denote that σ_i^2 is a positive function of $\boldsymbol{w}_i$.

From the definition, an ordered probit model is driven by an unobservable continuous random variable. The observed values, which have a natural ordering, can be regarded as categories representing the underlying process. See Figure 4 for a case of $k = 5$.

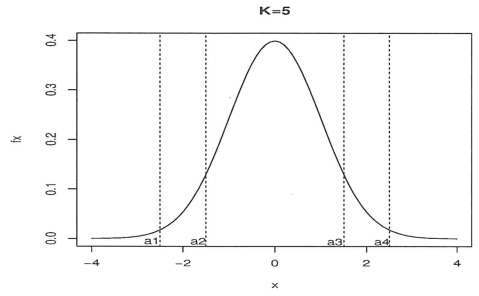

Figure 4: An illustration of ordered probit model with $k = 5$ categories.

The ordered probit model contains parameters $\boldsymbol{\beta}$, α_i $(i = 1, \ldots, k-1)$, and those in the conditional variance function $\sigma_i(\boldsymbol{w}_i)$ in Eq. (2). These parameters can be estimated by the maximum likelihood or Markov chain Monte Carlo methods. In this handout, we use the command `polr` of the R package `MASS` to estimate ordered probit models.

Example 6.1. To illustrate we consider the intraday price changes of Caterpillar stock on January 4, 2010. There are 37,716 transactions during the normal trading hours so that we have 37,715 price changes. For simplicity, we classify the price change into 7 categories shown in Table 1. Our analysis focuses on the dynamic dependence of intraday price changes. As such, we define indicator (or dummy)

Table 1: Frequencies of Price Change for Caterpillar Stock on January 4, 2010.

Category	1	2	3	4	5	6	7
Cents	< -2	$[-2, -1)$	$[-1, 0)$	0	(0,1]	(1,2]	> 2
Percentage	0.605	1.692	15.20	64.98	15.04	1.832	0.655

variables for lagged price changes:

$$y_{\ell,j} = \begin{cases} 1 & \text{if } y_{i-\ell} = s_j \\ 0 & \text{otherwise,} \end{cases}$$

where s_j denotes the jth category of price change and $y_{i-\ell}$ is the $(i-\ell)$th price change at time $t_{i-\ell}$, where $j = 2, \ldots, 7$ and $\ell = 1$ and 2. In other words, we employ the classifications of price changes for the previous 2 consecutive trades. As usual, with 7 categories, only six indicator variables are needed in modeling.

We also employ the observed price changes $y_{i-\ell}$ for $\ell = 1, 2, 3$ and the lag-2 transaction volume defined as $v_{i-2} = V_{i-2}/100$, where V_{i-2} is the actual volume. We do not use price volume because price is relatively stable in a trading day. Consequently, the model entertained is

$$\boldsymbol{x}_i\boldsymbol{\beta} = \beta_1 v_{i-2} + \sum_{\ell=1}^{3} \beta_{1+\ell} y_{i-\ell} + \sum_{j=2}^{7} \gamma_{1,j} y_{1,j} + \sum_{j=2}^{7} \gamma_{2,j} y_{2,j}. \quad (5)$$

For simplicity, we start with $\sigma_i^2(\boldsymbol{w}_i) = \sigma^2$, a constant. Parameter estimates of the model are given in Table 2, where all estimates but one are statistically significant at the usual 5% level. The parameter estimates of Eq. (5) are negative, because a negative sign is used in Equation (6). As a matter of fact, the model shown is a simplified one after removing some explanatory variables that were not statistical significant. For instance, we also included the time duration $\Delta t_i = t_i - t_{i-1}$ in the preliminary analysis and decided to drop the variable because its estimate is not statistical significant at the 5% level. The significance of the indicator variables shows that there exists dynamic dependence in intraday price change. The fitted model thus can be used to provide probability forecasts for the next transaction price change. Indeed, the model provides probability for each category of price change at each transaction.

indicator coeffts: $(\beta_1, \ldots, \beta_4)$

It is interesting to study the fitted boundary partitions of the ordered probit model in Table 2. First, because the explanatory variables may have nonzero means, the estimates of boundary parameters α_i are not symmetric with respect to zero. Second, $\hat{\alpha}_2 - \hat{\alpha}_1 = 0.577$ and $\hat{\alpha}_6 - \hat{\alpha}_5 = 0.601$. The two intervals roughly have the same length. Similarly, $\hat{\alpha}_3 - \hat{\alpha}_2 = 1.157$, which is close to $\hat{\alpha}_5 - \hat{\alpha}_4 = 1.140$. These results are consistent with the empirical observation that price changes appear to be roughly symmetric with respect to zero shown in Table 1.

Finally, the model implies

$$P(y_i^* \leq s_j | \boldsymbol{x}_i, \boldsymbol{w}_i) = \Phi\left(\frac{\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i}\right), \tag{6}$$

for the Caterpillar transaction data, where $\Phi(.)$ is the cumulative distribution function of $N(0,1)$.

Discussion The command `polr` allows for pre-determined weights to handle heteroscedasticity, but it cannot perform simultaneous estimation of the volatility and probit equations. See Hauseman, Lo, and MacKinlay (1992) and Tsay (2010) for some examples with time-varying $\sigma_i^2(\boldsymbol{w}_i)$ function. Finally, as usual, only 6 indicator variables are needed for each lagged value of y_i.

Table 2: Estimation Results of an Ordered Probit Model for the Intraday Price Changes of Caterpillar Stock on January 4, 2010 with 37,716 transactions. The Model is in Equation (5) and t Denotes t-ratio.

(a)	Boundary Partitions of the Probit Model					
Parameter	α_1	α_2	α_3	α_4	α_5	α_6
Estimate	−4.594	−4.017	−2.860	−0.853	0.287	0.888
t	−31.48	−27.80	−19.89	−5.944	2.000	6.188

(b)	Equation Prameters of Probit Model (estimates are negative)							
Par.	β_1	β_2	β_3	β_4	$\gamma_{1,2}$	$\gamma_{1,3}$	$\gamma_{1,4}$	$\gamma_{1,5}$
Est.	0.004	7.837	10.86	12.28	0.274	0.743	1.331	1.858
t	3.983	5.363	7.098	15.93	2.971	8.173	13.81	17.83
Par.	$\gamma_{1,6}$	$\gamma_{1,7}$	$\gamma_{2,2}$	$\gamma_{2,3}$	$\gamma_{2,4}$	$\gamma_{2,5}$	$\gamma_{2,6}$	$\gamma_{2,7}$
Est.	2.262	2.493	0.099	0.307	0.531	0.745	0.933	0.859
t	18.57	15.95	1.053	3.324	5.419	7.009	7.528	5.381

R Demonstrations for Ordered Probit Models

Output edited.

```
> da=read.table("taq-cat-t-jan042010.txt",header=T)
> head(da)
     date hour minute second price size

1 20100104    9     30      0 57.65 3910
 .....
6 20100104    9     30      1 57.65  462
> vol=da$size/100
> da1=read.table("taq-cat-cpch-jan042010.txt")
> cpch=da1[,1]   % category of price change
> pch=da1[,2]  % price change
> cf=as.factor(cpch)  % create categories in R
> length(cf)
[1] 37715

> y=cf[4:37715]
> y1=cf[3:37714]  % create indicator variables for lag-1 cpch
> y2=cf[2:37713]  % create indicator variables for lag-2 cpch

> vol=vol[2:37716]
> v2=vol[2:37713] % create lag-2 volume

> cp1=pch[3:37714] % select lagged price changes
> cp2=pch[2:37713]; cp3=pch[1:37712]

> library(MASS) % load package
> m1=polr(y~v2+cp1+cp2+cp3+y1+y2,method="probit")
> summary(m1)
Call:
polr(formula = y ~ v2 + cp1 + cp2 + cp3 + y1 + y2, method = "probit")

Coefficients:
       Value     Std. Error   t value
v2   -0.003765   0.0009453   -3.983
cp1  -7.836883   1.4613047   -5.363
cp2 -10.864394   1.5306456   -7.098
cp3 -12.283682   0.7710955  -15.930
y12  -0.274407   0.0923566   -2.971
y13  -0.742792   0.0908854   -8.173
y14  -1.330665   0.0963540  -13.810
y15  -1.858199   0.1042257  -17.829
y16  -2.261587   0.1218013  -18.568
y17  -2.493321   0.1563177  -15.950
y22  -0.098542   0.0935908   -1.053
y23  -0.307034   0.0923725   -3.324
y24  -0.531115   0.0980150   -5.419
y25  -0.744706   0.1062435   -7.009
y26  -0.932655   0.1238918   -7.528
y27  -0.858858   0.1596219   -5.381

Intercepts:
      Value    Std.Error  t value
1|2  -4.5941   0.1459    -31.4803
2|3  -4.0170   0.1445    -27.7989
3|4  -2.8599   0.1438    -19.8926
4|5  -0.8528   0.1435     -5.9437
5|6   0.2868   0.1434      1.9996
6|7   0.8882   0.1435      6.1883

Residual Deviance: 74802.56
AIC: 74846.56
> names(m1)
 [1] "coefficients"  "zeta"          "deviance"      "fitted.values"
 [5] "lev"           "terms"         "df.residual"   "edf"
 [9] "n"             "nobs"          "call"          "method"
[13] "convergence"   "niter"         "lp"            "model"
[17] "contrasts"     "xlevels"
> yhat=m1$fitted.values
> print(yhat[1:5,],digits=3)
        1       2       3       4      5      6      7
1 1.11e-03 0.005420 0.08605 0.660 0.2134 0.0266 0.007696
2 1.55e-02 0.041461 0.27883 0.608 0.0535 0.0028 0.000444
3 8.99e-06 0.000094 0.00522 0.287 0.4311 0.1605 0.116298
4 1.87e-04 0.001251 0.03267 0.539 0.3343 0.0658 0.027144
5 6.41e-04 0.003470 0.06457 0.630 0.2527 0.0365 0.011836
```