UNIST
School of Business Administration

FIN552 : High Frequency Financial Data Analysis — Daejin Kim, Spring 2017

Final Project
Due Date : **Monday, July 3, 2017**

1. Preliminary

1.1. Sampling

- The excel file **SP500_History.xlsx** contains a list of selected stocks consisting of S&P 500 index during 6-year (from 2003-09-10 to 2009-09-09).
- The excel file contains ticker information (same as *symbol_root* in TAQ file), listed stock exchange (N: NYSE, Q: NASDAQ), and company name.
- In the last two columns(Student ID, Student Name), **four stocks are assigned to each student** (We have 14 students, so we analyze 56 different stocks).
- **Do not download the TAQ data yet**!! Read carefully remaining instructions before starting analysis. Once reading carefully below instructions, select data with symbol_root = "your chosen ticker symbol" and symbol_suffix = "" (which is null) when you are ready.

1.2. Your tasks

- The first part in this project requires you to estimate or compute some variables at daily base. That is, your job is to construct daily time series. Second part requires you to perform some time-series analysis based on the data created in the part 1.
- During previous homework, you already did some jobs based on a particular date. This project will require you to do similar jobs for multiple days.
- For example, we can estimate effective spread of Apple stock on 2003-09-10, 2003-09-11, 2003-09-12, 2003-09-15, $\cdots$, 2009-09-09. (Remember that dates are not consecutive because of holidays, i.e. 2003-9-13 and 2013-9-14 are holidays).

1.3. Important steps for your project

- Do not download whole 6-year data at once. Downloading whole 6-year data requires lots of time and resources although we work on four stocks.
- Here are some broad outline to do this project:
  1. Download only two or three days of TAQ data for your assigned stocks.
  2. Write codes for each question to compute some variables.
  3. Test thoroughly, check whether there are any errors or not.
  4. Once you believe that your codes are correct, run your codes for entire sample by using appropriate looping commands.

1.4. Running whole data after checking your codes

- Running your codes for whole period (from 2003-09-10 to 2009-09-09) needs lots of computing time and resources. So, you need to do jobs into several steps. First, write your codes with small sample and test. Then, run your codes with entire sample. I think there are broadly two ways (local computing or cloud computing). Here is a local computing method.
  1. Download two-day of the TAQ trade, quote, and NBBO files for your assigned stocks.
  2. Write the first code to construct "BuySellIndicators" for each day and verify. (for e.g., the file name might be buysell20030910, buysell20030911, ...)
  3. Implement this code by using the PC-SAS, download all "BuySellIndicators" files from the WRDS into your PC.
  4. Write the second code to answer the questions. (the file name might be estim20030910, estim20030911, ...).
  5. *In sum, this procedure will run your codes in your local PC by using the downloaded "BuySellIndicators" files.*

  This process might be easy but still require lots of resources. The below is the second suggestion by using the WRDS-cloud service.

  1. Download two-day of the TAQ trade, quote, and NBBO files for your assigned stocks.
  2. Write one code (the first code + the second code in previous tip) to construct final outputs directly. That is, write code in order to create estim20030910, estim20030911, ... without downloading buysell20030910, buysell20030911, ...

3. Modify this code by using the PC-SAS, download all estim20030910, estim20030911, ... from the WRDS into your PC.
4. *In sum, this procedure will run your codes in the WRDS server without downloading the "BuySellIndicators" files.*

We have still some problems. That is, we need to run this code about 1,500 times (Step 3 or Step 4 in the first method or Step 3 in the second method) because we need to estimate on day 2003-09-10, 2003-09-11, and so on. This iterated job can be achieved by using looping statement (do - while in the SAS). If you have enough time, just loop about 1,500 times. But, here is a much simpler way to reduce your time. That is, use **qsas** in the WRDS-cloud.

1. After checking your codes in the local PC, you need to add some looping statements to handle multiple days.
2. Here is a tip. Replicate your codes : within some period, each code contains the same information but a running period is different. That is, job1.sas contains period from 2003-09-10 to 2004-09-09, job2.sas contains period from 2004-09-10 to 2005-09-09, and so on.
3. Go to the WRDS server, and type "**qsas job1.sas**" and "textbfqsas job2.sas", and so on. You can send this command at the same time without waiting the completion of your previous job.
4. Then, you get the same results but reduce the time.
5. Note that "proc download" will not be worked under the cloud environment. So, you need to modify your code not downloading your final files but storing your final files on some WRDS folders. (either in your home directory or /scratch/unist/...).
6. After finishing, download all files manually from the WRDS server into your local PC.

Now, are your ready? If you do not fully understand, read the WRDS cloud manual and this manual again.

2. Questions

1. For your analysis, this question will require you to construct some daily variables. Before start, let t denotes the original time variable in TAQ files (*time_m* in TAQ) and d denotes a day (2003-09-10, 2003-09-11, ...). Further, we use the following notations to represent variables.

   - p_t : trade price at time t.
   - r_t : stock return at time t, $r_t = \log(p_t) - \log(p_{t-1})$.
   - V_t : trading volume (*size* in trade file of the TAQ)
   - q_t : trading direction from Lee and Ready method
   - m_t : mid-quote point at time t.

   This exercise requires you to construct daily time series from intra-day data. That is, you need to construct the following table for each stock

Day (d)	p_d	m_d	V_d	DV_d	y_d	z_d	σ_d	A_d	s_d	es_d	res_d	λ_d^i	ψ_d^i	ρ_d	γ_d
030910	p_1	m_1	V_1	DV_1	y_1	z_1	σ_1	A_1	s_1	es_1	res_1	λ_1^i	ψ_1^i	ρ_1	γ_1
030911	p_2	m_2	V_2	DV_2	y_2	z_2	σ_2	A_2	s_2	es_2	res_2	λ_2^i	ψ_2^i	ρ_2	γ_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
090909	p_D	m_D	V_d	DV_D	y_D	z_D	σ_D	A_D	s_D	es_D	res_D	λ_D^i	ψ_D^i	ρ_D	γ_D

   (a) For each day d (from 2003-09-10 to 2009-09-09), identify the last closing price p_d and the last mid-quote point m_d. (Sometimes, you might not observe p_t for particular day due to no trading. In this case, you can treat p_d as missing.)

   (b) For each day d (from 2003-09-10 to 2009-09-09), compute daily share trading volume V_d and dollar trading volume DV_d. Further, define two order imbalances y_t and z_t at each time t and compute daily order imbalances (y_d and z_d). In sum,

compute the following daily variables

$$\begin{aligned} V_d &= \sum_{t=1}^{d_T} V_t \\ DV_d &= \sum_{t=1}^{d_T} V_t \times p_t \\ y_t &= q_t V_t - q_{t-1} V_{t-1} \\ y_d &= \sum_{t=1}^{d_T} y_t \\ z_t &= q_t V_t p_t - q_{t-1} V_{t-1} p_{t-1} \\ z_d &= \sum_{t=1}^{d_T} z_t \end{aligned}$$

where d_T is number of trades at day d.

(c) For each day d (from 2003-09-10 to 2009-09-09), compute the intra-day volatility (σ_d) of return series which is the standard deviation of return series r_t.

(d) For each day d (from 2003-09-10 to 2009-09-09), compute the Roll's implicit spread measures, i.e., $s_d = \frac{1}{2}\sqrt{-Cov(\Delta p_t, \Delta p_{t-1})}$ for each stock. Between 2003-09-10 and 2009-09-09, plot the s_d for each stock. Your graphs should have day on the horizontal axis, and s_d on the vertical.

(e) Amihud(2002) defines a simple version of illiquidity measure. His measure is originally developed for computing monthly measure based on the daily data such as CRSP. In this question, we try to compute the daily Amihud' measure from intra-day data. For each day d (from 2003-09-10 to 2009-09-09), Compute Amihud's measure (A_d)from your data where A_d is defined as

$$A_d = \sum_{t=1}^{d_T} \frac{\text{Absolute Value of } r_t}{p_t \times V_t}$$

where d_T is number of trades at day d.

(f) For each day d (from 2003-09-10 to 2009-09-09), compute the daily volume(size) weighted average of effective spread (es_d) and relative effective spread (res_d) for each stock where $res_t = es_t/m_t$. That is $es_d = 1/V_t \sum_{t=1}^{T} V_t(es_t)$ and $res_d = 1/V_t \sum_{t=1}^{T} V_t(res_t)$. Between 2003-09-10 and 2009-09-09, plot the es_d and res_d for each stock. Your graphs should have day on the horizontal axis, and es_d on the vertical.

(g) For each day d (from 2003-09-10 to 2009-09-09), estimate the following two regressions for each stock.

$$\begin{aligned} r_t &= \alpha_d^1 + \lambda_d^1 y_t + \psi_d^1 y_{t-1} + \varepsilon_t \\ r_t &= \alpha_d^2 + \lambda_d^2 z_t + \psi_d^2 z_{t-1} + \varepsilon_t \end{aligned}$$

where λ_d^i is price impact coefficient and ψ_d^i is price reversal coefficient on day d with $i = 1, 2$. Between 2003-09-10 and 2009-09-09, plot the λ_d^i for each stock. Again, plot the ψ_d^i for each stock. Your graphs should have day on the horizontal axis, and λ_d^i or ψ_d^i on the vertical.

(h) The quote direction should be identified from intra-day data. Handling intra-day data, however, requires lots of time and efforts. Thus, many researchers prefer modeling the quote direction either $AR(p)$ model or $MA(q)$ model. For each day d, fit the following $AR(1)$ and $MA(1)$ model to estimate coefficients ρ_d and γ_d.

$$\begin{aligned} AR(1) : q_t &= \rho_d q_{t-1} + v_t \\ MA(1) : q_t &= v_t + \gamma_d v_{t-1} \end{aligned}$$

2. Based on your constructed daily sample in question 1, perform the following analysis.

   (a) Provide correlation table for variables $\sigma_d, A_d, s_d, es_d, res_d, \lambda_d^1, \psi_d^1, \lambda_d^2, \psi_d^2, \rho_d, \gamma_d$.

   (b) Define two return series $r_d = \log(p_d/p_{d-1})$ and $w_t = \log(m_d/m_{d-1})$ from daily closing price(p_d) and midpoint price(m_d). Fit the following four regressions separately for each stock

$$\begin{aligned} r_d &= a_1 + b_1 y_d + b_2 y_{d-1} + \varepsilon_t \\ r_d &= a_1 + b_1 z_d + b_2 z_{d-1} + \varepsilon_t \\ w_d &= a_1 + b_1 y_d + b_2 y_{d-1} + \varepsilon_t \\ w_d &= a_1 + b_1 z_d + b_2 z_{d-1} + \varepsilon_t \end{aligned}$$

   (c) Run the following regressions for each stock

$$\begin{aligned} s_d &= a_1 + b_1 \log(DV_d) + b_2 \sigma_d + b_3 \log(p_d) + \varepsilon_t \\ es_d &= a_1 + b_1 \log(DV_d) + b_2 \sigma_d + b_3 \log(p_d) + \varepsilon_t \\ res_d &= a_1 + b_1 \log(DV_d) + b_2 \sigma_d + b_3 \log(p_d) + \varepsilon_t \end{aligned}$$