Introduction to SAS, CRSP, Compustat and TAQ[1]

Instructor: Dr. Yuxing Yan
Email: yuxing.yan@canisius.edu
yuxing.yan@faculty.umuc.edu

Objectives:

1) Learn how to use SAS to conduct empirical research in finance/accounting areas
2) Learn and use CRSP database which contains trading data for all stocks listed in the US from 1926 onward
3) Learn and use Compustat (Research Insight, Capital IQ) database which contains accounting information such as balance sheet, income statement for all public firms in the US from 1950s onward
4) Learn and use TAQ (Trade and Quote) database which contains high-frequency (second by second) transaction data for all publicly traded stocks in the US from 1993 onward

Target students:

1) Graduate students who plan to apply for a Ph.D. program in finance or accounting
2) First-year doctoral students who intend to do empirical research in finance/accounting
3) Researchers/students who are interested in big data analytics especially in the areas of finance or accounting

Required textbooks

1) Delwiche, Lora D. and Susan J. Slaughter, Little SAS book: A primer, *SAS Institute Inc.*, 4th edition
2) Böhmer, Ekkehart, John P. Broussard and Juha-Pekka Kallunki, 2002, Using SAS in Financial Research, *SAS Institute*, ISBN-10: 1590470397, ISBN-13: 978-1590470398
3) Stock Market Analysis using the SAS System: Technical Analysis, SAS Institute Inc., ISBN 1-55544-222-6.
4) Lecture notes

Prerequisites:

1) Basic finance courses covering corporate finance and/or investment
2) Your school must have valid CRSP and Research Insight/Compustat/Capital subscriptions.

Software: SAS (PC SAS or UNIX SAS)
Teaching Method: online via email, phone, Skype etc.
Web sites: (add later)
Efforts expected: minimum 2 hours pay day including weekends
Duration of the course: 14 weeks
Focus of the course: hands-on

[1] 5/17/2013 Yuxing Yan

Grading Policy:

Homework (14)	30%
Data cases (10)	40%
Group project	20%
Asking questions	10%
Total	100%

Group project

Each group could have up to two members. A topic should be closely associated with this course. The maximum of your will be 15, one-sided, double spaced, font of 12. Please discuss with me your topic before you start to work on it.

Real world topics are especially encouraged. Three parts are essential:

1) theory and background of the topic,
2) SAS programs and an explanation of the codes,
3) final data set plus the codes to process the data

Note: see a list of potential topics at the end of this document

Tentative schedule

Week 1:

SAS	input data, text input file, csv input file, PC SAS vs. UNIX SAS, program window, log window and output winder, header file
Data	1) from DS[2] 2) Fama-French monthly factors http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
Readings	1) DS chapters 1 and 2 2) Fama, Eugene and Kenneth R. French, 1992, The cross-section of expected stock returns, Journal of Finance 47, 427-465.
Assignments	1) HW #1 (about 10 questions) 2) Download and generate SAS Fama-French monthly factors

[2] The first text book by Delwiche and Slaughter.

Week 2:

SAS	Read from an existing SAS data set, libname, proc contents
Data	1) from DS 2) CRSP: stockInfo.sas7bdat Definitions: PERMNO, PERMCO, ticker, begdate enddate, CUSIP
Readings	1) DS chapter 2 2) CRSP manual
Assignments	1) HW #2 2) generate the 9th digit for a given CUSIP

Week 3:

SAS	Output data: text file, csv file or SAS data set How debug your program? Title, seeing is believing
Econometrics	F-test, T-test, p-values
Data	1) from DS 2) CRSP: indexMonthly, indexDaily SAS data sets Definitions: EVRETD, VWRETD
Readings	1) DS : chapter 3 working with your data 2) BBK: chapter 1
Assignments	1) HW #3 2) Data Case #1: January effect/weekday effect

Week 4:

SAS	Date format, real, integer, year, month functions Proc sort, merge different data sets Remove unnecessary SAS data sets
Econometrics	F-test, T-test, p-values
Data	CRSP: misx.sas7bdat, decile data
Readings	1) DS: chapter 4 sorting 2) BBK:[3]
Assignments	1) HW #4 2) Data Case #2: which party manages stock market better? Merge index and stock monthly

[3] The second textbook by Bohmer, Broussard and Kallunki.

Week 5:

SAS	Advanced proc sql;
Data	CRSP stockMonthly SAS data set
Readings	1) DS 2) BBK: chapter 3 Analysis winners and losers 3) Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: Implications for stock market efficiency, Journal of Finance 48, 65-91.
Assignments	1) HW #5 2) Data Case #3: Momentum strategy

Week 6:

SAS	proc means, transpose
Data	CRSP: stockDaily, EWRETD, EWRETD Output trading data for a given set of tickers (PERMNO) for a text file for Excel usage, merge index with stock
Readings	1) DS chapter 5 combining SAS data sets 2) BBK
Assignments	1) HW #6 2) data case #4: replicate S&P500 index 3) data case #5: replicate CRSP VWRETD, EWRETD

Week 7:

SAS	First and last function, Retain
Data	CRSP eventMonthlyEvent SAS data set
Reading	1) BBK 2) Scholes, Myron and Joseph Williams, 1977, Estimating Betas from Nonsynchronous Data, Journal of Financial Economics, 5, 309-327.
Assignments	1) HW #7 2) How to select stocks listed on NYSE and AMEX? 3) data case #6: beta estimation

Week 8:

SAS	Rank groups, Proc sql;
Data	CRSP: decile returns based on size Research Insight: AnnualFin
Reading	1) DS chapter 6 SAS macro 2) BBK 3) Roll, Richard, 1984, A simple implicit measure of the effective bid-ask spread in an efficient market, Journal of Finance 39, 1127-1139.
Assignments	1) HW #8 2) Data Case #7: replicate CRSP decile return

Week 9:

SAS	Speed issues, remove unnecessary data sets Scan function
Data	CRSP: eventDaily SAS data sets, NYSE/AMEX stocks only
reading	1) DS 2) BBK: 3) Amihud, Y., 2002. Illiquidity and stock returns: Cross-section and time-series effects, Journal of Financial Markets 5, 31–56.
Assignments	1) HW #9 2) Data Case #8: illiquidity measure (Amihud, 2002)

Week 10:

SAS	Proc sql, simple macro variable
Data	Compustat, CRSP: historical vs. header ticker Historical CUSIP vs. header CUSIP
Readings	1) DS 2) BBK
Assignments	1) HW #2 2) Merge CRSP with Compustat 3) Data Case #9: Rolling Beta estimation

Week 11:

SAS	function (macro function)
Data	CRSP daily data
Reading	1) DS 2) Pastor, L. and R. Stambaugh, 2003, Liquidity risk and expected stock returns. Journal of Political Economy 111, 642-685.
Assignments	1) HW #11 2) Data Case #10: liquidity measure by Pastor and Stambauch,2003

Week 12:

SAS	Speed issue More on macro
Data	TAQ High-frequency data: CT (consolidated trades) , CQ (consolidated quotes)
Readings	BBK chapter 10 Analysis of transaction cost
Assignments	1) HW #12 2) Data Case #11: estimate spread from CQ

Week 13:

SAS	Generate macro variables by using proc sql
Data	1) TAQ
Reading	1) BBK chapter 10 Analysis of transaction cost
Assignments	1) HW #13 2) Merging CT with CQ

Week 14:

SAS	When log file is too big
Data	TAQ
readings	1) BBK chapter 10 Analysis of transaction cost 2) Lee, C. and M. Ready, 1991, Inferring Trade Direction from Intraday Data, Journal of Finance, 46, 2, 733–746.
Assignments	1) HW #14 2) Data case #12: trading direction, Lee and Ready (1991)

Topics for a term project

Potential topic #1 : PIN estimation[4]

Data	High-frequency data: CT (consolidated trades) , CQ (consolidated quotes) Hasbrouck's TORQ database
readings	Easley, David, Nicholas M. Kiefer, and Maureen O'Hara, and Joseph B. Paperman,1996, Liquidity, information, and infrequently traded stocks, Journal of Finance 51,1405–1436.

Potential topic #2: Replication of Fama-French factors

Data	High-frequency data: CT (consolidated trades) , CQ (consolidated quotes) Hasbrouck's TORQ database
Readings	Fama, Eugene and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, Journal of Financial Economics 33, 3056. Fama, Eugene and Kenneth R. French, 1992, The cross-section of expected stock returns, Journal of Finance 47, 427-465.

Potential topic #3: KMV default probability

Data	1) hypothetical data 2) CRSP and Compustat
Readings	Merton, Robert C.,1973,,Theory of Rational Option Pricing, *Bell Journal of Economics and Management Science* (The RAND Corporation), **4** (1): 141–183.

Potential topic #4: Spread from high and low daily prices

Data	CRSP daily data
Readings	Corwin, Shane and Paul Schultz, 2012, A simple way to estimate bid-ask spreads from daily high and low prices, Journal of Finance 67(2), 719-759.

[4] Extremely difficult.

Potential topic #5: Trading direction (Ellis et al, 2000)

Data	TAQ
readings	1) BBK chapter 10 Analysis of transaction cost .2) Ellis, K., Michaely, R., O'Hara, M., 2000. The accuracy of trade classification rules: evidence from Nasdaq, Journal of Financial and Quantitative Analysis 35, 529–551.

Potential topic #6: misclassification of trading directions, Boehmer et al. (2007)

Data	TORQ
Readings	1) TORQ manual 2) BBK chapter 10 Analysis of transaction cost 3) Boehmer, E., Grammig, J., Theissen, E., 2007. Estimating the probability of informed trading -- Does trade misclassification matter? Journal of Financial Markets 10, 26 47.