Paper 202-2009

SAS® in Financial Research: Embracing Data

Taras Zlupko, CRSP®, University of Chicago, Chicago, IL

ABSTRACT

Excellent introduction to some of the ideas of empirical financial economics and mechanics of their technical exploration and investigation with SAS is provided by Boehmer et al. in "Using SAS in Financial Research". The book makes extensive use of CRSP and Compustat which are some of the most reputable databases available to financial researcher. It is also assumed that the data required by each particular study is pre-processed and resides in external files ready to be exported into SAS for further analysis. This implies involvement of some tools outside of SAS data access.

This paper shows how to directly engage the empirical data into financial research with SAS without needing other media. This is primarily based on the use the SESECRSP engine. The engine provides convenient and elegant way of integrating data access and preparation into the process of implementing research plan with SAS by presenting CRSP and Compustat databases as native SAS data sets. Bringing together data access, preparation and analytical strengths of SAS contributes to higher efficiency of quantitative financial research investigation.

INTRODUCTION

"Using SAS in Financial Research" by Boehmer et al. provides insightful discussion of some of the most popular research topics in finance such as random walk theory, overreaction hypothesis, the capital asset pricing model, price-earnings relationship, forecasting market performance, event studies and others. Ease and fluency of SAS in implementing the methodology of different types of empirical research in finance comes out naturally throughout the course of the book.

Availability of and access to quality data is a major factor impacting any empirical study. "Using SAS in Financial Research" mainly makes use of CRSP and Compustat which are broadly used by financial research community. It is an assumption in the book that the data required to carry out a particular research using CRSP and Compustat is pre-arranged into external files which are ready to be exported into SAS for further processing and analysis. This implies some intermediate steps and media employed in order to access the original databases and handle initial data preparation. Being not necessarily straightforward and done outside of SAS, this makes the case about having SAS working directly with the aforementioned databases for both analysis as well as data access and preparation in uninterrupted flow.

CRSP stock, indices, CRSP/Compustat Merged database aw well as the Compustat database available through CRSP are provided in the proprietary CRSPAccess format. This paper illustrates how to directly access the data provided by CRSP in its proprietary format into financial research process with SAS. This is mainly done through the use of the SESECRSP engine which was introduced after "Using SAS in Financial Research" was published. Accessing empirical data directly from SAS allows for additional flexibility in testing potential extensions of research methodology and enriches overall research experience.

VARIANCE RATIO TESTING

The variance ratio test of the random walk hypothesis developed by Lo and MacKinlay (1988) uses weekly values of CRSP equal- and value-weighted NYSE/AMEX market index[1]. The data is selected on Wednesday of each week from the daily CRSP file. If Wednesday falls on a non-trading day, then Thursday's value is used for that week. If both Wednesday and Thursday are non-trading days of the week, then Tuesday is selected to represent the week.

The data required for this study is coming from CRSP Indices database. The indices product is available in both CRSPAccess and SAS format from CRSP. Therefore the data can be accessed directly from SAS with and without making use of the SASECRSP engine. Code below shows how this can be accomplished both ways using CRSP NYSE/Alternext market index equal-weighted index.

[1] CRSP NYSE/AMEX index has been renamed to the CRSP NYSE/Alternext index in October 2008 following acquisition of the American Stock Exchange by NYSE Euronext.

```
/*1.  Create SAS dataset from CRSPAccess database*/

/*1.1 Set up library restricting data to market index of interest
      (daily CRSP NYSE/Alternext equal-weighted index) and
      to period from September 05, 1962 to December 27, 1985*/

libname ind SASECRSP 'C:\crspdata\dailydata\'
                          setid=460
                             indno=1000041
                                 range='19620905-19851226';


/*1.2 Select price index values on Tuesdays, Wednesdays and Thursdays and
      convert CRSP date to SAS date*/

data i; set ind.aind(drop=indno);
             caldt=crspdcsd(caldt);
                    wday=weekday(caldt);
                           if wday in (3,4,5);
run;


/*2.  Create the same dataset as produced in 1. above from CRSP
      standalone index product*/

/*2.1 Set up library*/

libname ind 'C:\crspdata\indexdata\';

/*2.2 Select price index values on Tuesdays, Wednesdays and Thursdays.
      No need to convert dates as they are provided in SAS format.*/

data i; set ind.dsic(keep=caldt ewindx) ;
             where '05Sep1962'd<=caldt<='26Dec1985'd;
                    wday=weekday(caldt);
                           if wday in (3,4,5);
run;


/*3. Create week identifier following either 1. or 2. above and
     prioritize days of week according to research design*/

data i; set i;
    retain w;
      if _n_=1 then w=1;
                else do; if caldt-lag(caldt)<3 then w=w; else w=w+1; end;
                     if wday = 4 then wday=1;
                     if wday = 5 then wday=2;
run;


/*4. Select best available value by sorting data by priority day each
     week and keeping only top observation*/

proc sort data=i out=i; by w wday;
data i; set i;
             by w;
                  if first.w=1;
run;
```

As seen above, having access to data available from CRSP in either SAS format or in the CRSPAccess format through the SASECRSP engine allows incorporating data preparation step into a single workflow of research investigation within SAS environment. In addition to accommodating the requirements of research methodology this makes it possible to conveniently update empirical data inputs should modifications to research design require this. For example, it is fast and easy to update the empirical data used in the test based on new requirements related to time period of interest or by including another CRSP index series into analysis, such as NYSE/Alternext value-weighted index which is also used in Lo and MacKinlay (1988).

Time series of weekly CRSP NYSE/Alternext equal-weighted index values which have been generated from the daily index in accordance with the variance testing specification allows to immediately proceed with further steps of the research. This includes obtaining log-relative changes of the index. One week log-relative changes can be calculated as follows:

```
/*5. Weekly log-relative changes of the NYSE/Alternext equal-weighted
     market index*/

data i; set i;
                lrc=log(ewindx)-lag(log(ewindx));
run;
```

Figure 1 below shows weekly log-relative changes of the CRSP NYSE/Alternext equal-weighted market index.

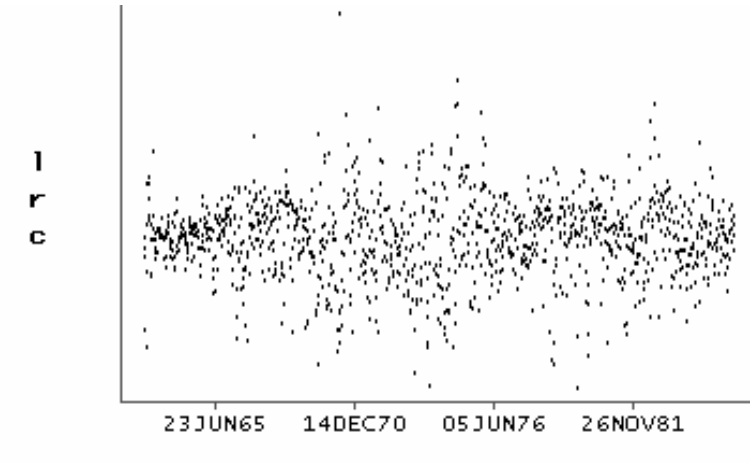

Figure 1. Log-relative weekly changes of the CRSP NYSE/Alternext equal-weighted market index, September 1962 – December, 1985.

The variance ratio testing specification requires creating additional series of log-relative index changes based on two-, four-, eight- and sixteen- week periods. Generating these series and completing further steps of the variance ratio testing can be done as presented in “Using SAS in Financial Research”.

Therefore both data preparation steps and data analysis for variance ratio testing can be conveniently done within SAS environment. Incorporating direct data access into a single process of empirical financial testing using SAS tools adds flexibility and efficiency to the testing process.

OVERREACTION HYPOTHESIS

As in some other chapters of “Using SAS in Financial Research”, the assumption is made that input data set is readily available and is generated by some process and tools outside of SAS. Therefore SAS is employed from the point of reading the external data set and then proceeding further steps. Since data used in the overreaction hypothesis testing comes from CRSP, having the SASECRSP engine in place makes it possible to implement full cycle research entirely within SAS without recourse to other media. This is particularly useful given the well-known SAS strengths in preparing and working with data sets. As shown

below, data to test the overreaction hypothesis can be conveniently and transparently created with SAS accessing CRSP stock and indices database directly.

For the purpose of illustration we will show how to access monthly security returns in CRSP database. These then can be used for further processing in order to perform the overreaction hypothesis testing. Index returns can be accessed as shown in the earlier chapter on the variance ratio testing.

The basic idea is to obtain total returns data at security level. These returns then are used to calculate cumulative excess returns (i.e. market-adjusted returns)[2] for each security and to select best and worst performing stocks. The period is usually set to three years. The stocks selected are then included in two separate portfolios of past winners and losers respectively and the portfolios are evaluated over the subsequent three year period. The process of selecting top and bottom performing stocks and following related portfolios afterwards is repeated over every three year subsequent non-overlapping periods during some long-term historical time range of interest. Therefore each 'portfolio formation' period is followed by portfolio evaluation period of the same length. The performance of winners and losers during the evaluation period is of interest only at the portfolio level.

There are different ways to proceed with creating the dataset for testing the overreaction hypothesis using CRSP data. However, in essence the task is to show that CRSP returns data is accessible directly from SAS. The rest of the processing could be done according to preferences of individual researcher performing the test and particulars of test design. CRSP returns are hard coded in the database and will be available in a separate table as soon as the libname statement using SASECRSP engine is executed.

```
/*6. Access CRSP monthly stock database*/

libname mstk SASECRSP 'C:\crspdata\monthlydata\' setid=20;
```

Please note that in addition to pointing to a specific physical location of monthly CRSP database, the libname statement also needs to use a database-specific SETID. The latter is reflective of the type of database being accessed.

If process design would benefit from having database restricted to a particular date range this can be done using range option. For example, De Bondt and Thaler (1985) use the returns from January 1930 to December 1932 as their first three-year 'portfolio formation' period and repeat the process sixteen times with last three-year portfolio formation period starting in January 1975 and ending in December 1977. Given the above and the required three-year 'portfolio evaluation' period, the date range available to researcher will need to be set from January 1930 to December 1980. Restricting the database to specific period of interest may help in speeding up further processing as SAS would not need to read through the entire database. This can be done using SASECRSP range option as follows:

```
/*7. Access CRSP monthly stock database with time restriction*/

libname mstk SASECRSP 'C:\crspdata\monthlydata\' setid=20
                                        range='19300101-19801231';
```

In order to add the added speed of execution, the entire process of determining winners and losers within each portfolio formation period as well as calculating portfolio returns may be designed within a framework of performing these iterative steps on a subset of database restricted to the period of given iteration. For example if CRSP database is already restricted to three years by relevant libname statement then time will be saved each time the time bound selection of data is performed. Restricting the database by the range option is done momentarily.

```
/*8. Iterative libname assignment with macro*/

%macro port_formation;
data _nill_;
      %do i=30 %to 77 %by 3;
            libname mstk SASECRSP 'C:\crspdata\monthlydata\'
                        setid=20
                        range="19%eval(&i)0101-19%eval(&i+2)1231";
```

[2] Please see Conrad and Kaul (1993) for additional discussion.

```
        /*determine winners and losers for given portfolio formation
        period, save results then move on to next portfolio formation
        period (not shown)*/
      %end;
%mend;
%port_formation;
```

The code above provides example of using libname statement within SAS macro. New three-year database is set up during each iteration and further processing can be done on the entire restricted database (not shown in the code above) rather then passing through the entire database when selecting data subset for a given period of interest.

EVENT STUDIES

Event studies are focused on investigating the impact of new information about the company on its stock returns. Date when the information becomes available is usually the event date. Therefore differences in returns before and after the event date are contrasted. Usually different types of events are investigated separately. “Using SAS in Financial Research” provides example of event study based on earning surprises. We will use stock splits for the purpose of illustration. Stock splits were used in the pioneering work in the area of event study methodology by Fama et al. (1969). As much as splits are intuitively appealing for the purpose of this type of research, they also lend themselves well in illustrating the methodology.

Ultimately same type of dataset is needed for any event study whether focusing on earning surprises, stock splits etc. The dataset would need to include returns history before and after the event. These are called estimation and event periods. The first task is thus to identify event of interest and stocks affected by the event. Then make sure there is sufficient trading history for each stock to be included in the study. Therefore, event study analyzing stock splits would need dates of split announcement first. This can be obtained directly from CRSP database using CRSP distribution code 5523 which is assigned to split events.

```
/*9.  Create dataset for event study*/

/*9.1 Set up SAS library accessing CRSP daily stock database restricted
      to last ten years*/

libname dstk SASECRSP 'C:\crspdata\dailydata\' setid=20
                                        range='19990101-20081231';

/*9.2 Select securities that announced a stock split between January 1,
      2000 and December 31, 2008, one (last) event per security;
       - assume 200-trading day history in used in estimation period and
         only day of announcement is used in evaluation period;
       - create date in crspdate format in order to be able working with
         trading not calendar dates*/;

data split(keep=permno dclrdt date); set dstk.dists;
      by permno; if last.permno;
           where distcd=5523 and
                '01Jan2000'd<dclrdt<'31Dec2008'd;
                     date=crspdscd(dclrdt,1);
run;

/*9.3 Partition into estimation and event period and add daily returns;
      estimation period is set at 190 trading days ending ten days before
      the event*/

data splitest splitevt;
      merge split(in=good) dstk.ret;
      by permno; if good=1;
            if date-200<=caldt<=date-11 then output splitest;
            if caldt=date then output splitevt;
run;
```

```
/*9.4 Keep only stocks with 190 days of trading history in estimation
      period*/

proc sql; create table temp as select permno, count(permno)
                from splitest group by permno having count(permno)=190;
           create table splitest as select a.* from splitest a, temp b
                where a.permno=b.permno;
           create table splitevt as select a.* from splitevt a, temp b
                where a.permno=b.permno;
quit;

/*9.5 Set up SAS library accessing CRSP daily indices database restricted
      to last ten years and to CRSP NYSE/Alternext/NASDAQ/Arca value-
      weighted market index only*/

libname dind SASECRSP 'C:\crspdata\dailydata\'
                           setid=460
                              indno=1000200
                                 range='19990101-20081231';
```

Please note that 9.1 and 9.5 access the same combined CRSP stock and indices database. The libraries representing different daily databases are assigned by using relevant SETID and are presented as SAS native data sets.

Having the indices database restricted to only one market index of interest allows proceeding with merging stock and market returns without defining specific index in the code as follows. Merging is done by date filed which is in the same CRSP date format in the datasets involved. Once security and market returns are merged, regression can be run to evaluate the relationship between stocks and the market.

```
/*9.6 Merge estimated period stock and market returns and run
       regressions*/

proc sql;
       create table splitest as select a.*, b.tret
                           from splitest a, dind.tret b
                                  where a.caldt=b.caldt;
       create table splitevt as select a.*, b.tret
                           from splitevt a, dind.tret b
                                  where a.caldt=b.caldt;
quit;

proc sort data=splitest; by permno caldt;

proc reg data=splitest
          outest=estparam(rename=(intercept=alpha tret=beta)
                                  keep=permno intercept tret) noprint;
              by permno;
                model ret=tret;
                 label tret=beta;
run;
```

Now abnormal returns on event date can be calculated for each stock using regression estimates.

```
/*9.7 Calculate abnormal returns of individual stocks*/

proc sort data=splitevt; by permno;

data splitevt; merge splitevt estparam;
              by permno;
                   abnret=ret-alpha-beta*tret;
run;
```

Abnormal returns of stocks on announcement day are plotted in Figure 2.

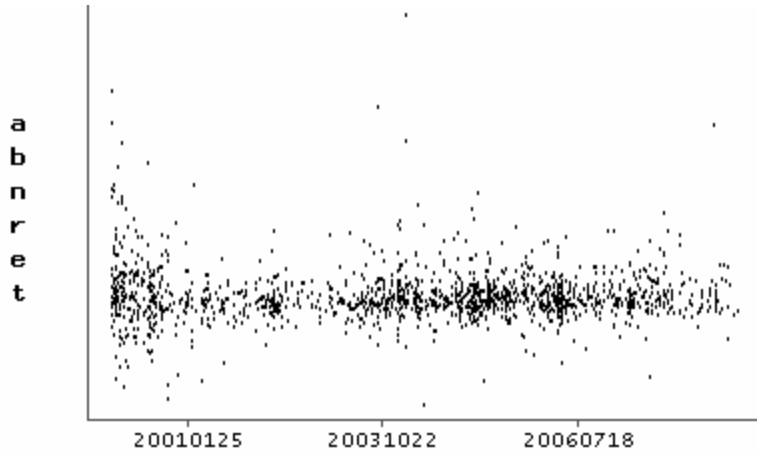

Figure 2. Abnormal returns of stocks on split announcement day.

Finally mean abnormal return on split announcement day can be evaluated for statistical significance.

```
/*9.7 Calculate mean abnormal return and evaluate significance*/

proc means data=splitevt n mean t;
                          var abnret;
run;
```

The above illustration of event study approach shows conceptual benefits of direct SAS connection with data used in the study. Actual progress of event study would depend on specific study goals and design and may include relevant checks and conditions. Since SAS access required data directly, all necessary tasks can be performed within SAS environment.

CONCLUSIONS

"Using SAS in Financial Research" by Boehmer et al. provides useful introduction to employing SAS when working on some popular topics in empirical finance such as testing random walk theory, overreaction hypothesis, the capital asset pricing model, event studies or price-earnings relationship and others. Conducting these studies would benefit from direct access to CRSP and Compustat data from SAS which is possible by using the SASECRSP engine.

Accessing CRSP and Compustat databases using the SASECRSP engine allows presenting these databases in native SAS format and engaging the power of SAS data processing and analytical tools to fuller extent. Availability of direct access to CRSP and Compustat databases from SAS allows increased flexibility and efficiency when conducting empirical financial research using these databases.

REFERENCES

Boehmer, E., Broussard, J.P. and J.-P. Kallunki. 2002. Using SAS in Financial Research. Cary, NC: SAS Institute Inc.

Conrad, J. and G.Kaul. 1993. Long-Term Market Overreaction or Biases in Computed Returns? *The Journal of Finance* 48: 39-63.

De Bondt, W.F.M. and R. Thaler. 1985. "Does the Stock Market Overreact?". *The Journal of Finance* 40: 793-805.

Fama, E.F., Fisher, L., Jensen, M.C. and M. Roll. 1969. "The Adjustment of Stock Prices to New Information". *International Economic Review* 10: 1-21.

Lo, A.W. and A.C. MacKinlay. 1988. “Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test”. *The Review of Financial Studies* 1: 41-66.

SAS Institute Inc. 2008. *SAS/ETS® 9.2: User’s Guide*. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/cdl/en/etsug/60372/PDF/default/etsug.pdf

Zlupko, Taras. 2006. “Compass Rose in Stock Markets: The Power of SASECRSP LIBNAME.” *Proceedings of the Thirty First Annual SAS Users Group International Conference*, San Francisco, CA, 183-31.

Zlupko, Taras. 2007. “CRSP Indices in SAS”. *Proceedings of SAS Global Forum*, Orlando, FL, 167-2007.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Taras Zlupko
CRSP, University of Chicago
105 West Adams Street, Suite 1700
Chicago, IL 60603
Work Phone: (312)263-6654
E-mail: Taras.Zlupko@crsp.ChicagoBooth.edu