WRDS & SAS INTRODUCTION

for PhD Students

AHMET TUNCEZ

5/20/2013

Contact email: atuncez@mays.tamu.edu

This document includes a collection of useful links; however, some of them may turn out to be broken links in the future. I thank Nan Yang for his help.

1. INTODUCTION

Wharton Research Data Services (WRDS) is the leading data platform for business research. WRDS is an internet-based system and all files are in SAS format.[1]

2. SAS BASICS

- As a starter, you can visit the UCLA website to find resources to help you learn and use SAS: http://www.ats.ucla.edu/stat/sas/
  - You can start with SAS Starter Kit and continue with Classes and Seminars and Learning Modules

- Also , you can check the official SAS website: http://www.sas.com/ and SAS Support pages: http://support.sas.com

- Some tutorials can be found from Texas A&M's Statistics department website: http://dist.stat.tamu.edu/flash/SAS/

- These two books are very helpful:

  Boehmer, E., Broussard, J. P., & Kallunki, J. P. (2002). Using SAS in financial research. SAS Institute.

  This book illustrates how to use SAS software to conduct basic empirical analyses of stock market and financial statement data.

  Delwiche, L. D., & Slaughter, S. J. (2012). The little SAS book: A primer. SAS Institute.

  You can find the most commonly-used features of SAS software with examples and graphics in this book.

[1] Using web query, you can easily import the data into other statistical packages including Stata or Excel.

3. SSH Basics

Why do you need SSH when using WRDS? Simply put, if you want to download data and work locally, you can use it. Also, you may want to work on a UNIX environment. Thus, you may use SSH either as a main tool in your study or a complementary tool for data transfers between local and WRDS systems.

You can connect the WRDS Unix server using the SSH Secure Shell. You can download SSC Clients from here.[2] You can create bookmarks using SSH Secure Shell.

To connect to WRDS:

Host name is *wrds.wharton.upenn.edu* and use your *user_name* and *password.*

4. WRDS Basics

WRDS website is https://wrds-web.wharton.upenn.edu/wrds/

Mays Business School provides access to WRDS, which only faculty/staff/graduate students in our school can legally access. If you would like an account, please go to http://wrds-web.wharton.upenn.edu/wrds/index.cfm?true and select the PhD option under register. When you sign up, please use an official A&M email address.[3]

After you sign up, you can login with your *user_name* and *password.*

There are three ways to use WRDS:

1. Web Query
2. Unix environment (SAS, Fortran)
3. PCSAS (SAS)

Note that you can download the same data using any of these methods.

[2] Also, you can check alternative SSH clients: http://sc.tamu.edu/help/general/accessMethods/sshclients.html

[3] If you already have an account but do not remember your username and password, do not request a new account. You can use the "forgot your username/password" link on the log in page.

i. Web Query (SAS, Stata, Excel, etc.)

This is the easiest way to explore data. To use web query, you just need a browser. You can choose the data items in the datasets and select among different output formats (for example, SAS, Stata, Excel, etc.). Once the output is ready, you may save it to your PC.

You can download up to 2 gigabytes of data per download.

ii. Unix environment (SAS, Fortran)

To use the Unix environment, you need a secure connection with SSH client. Unix code and programming in SAS are needed. Permanent disk space is 750 MB in the home directory (/tamu/user_name) and temporary disk space is 2TB for processing programs and storage.

iii. PCSAS (SAS)

You need SAS software installed on your desktop. You can obtain remote access to a data source by connecting to WRDS. Also, you can access your Unix disk space.

To remote sign on to the WRDS Server, you need to add this code at the beginning of your program:

```
/****************************************
Remote Sign-on to WRDS Server
****************************************/
%let wrds = wrds.wharton.upenn.edu 4016;
options comamid=TCP remote=WRDS;
signon username=_prompt_;
```

Also, you need to write your codes between these two codes:

```
rsubmit;

endrsubmit;
```

Note that, to download the data in your local drive, you may need this code
(before endrsubmit)

```
proc download data=crsp_june    out=mylocal.crsp_june;
run;
```

5. WRDS Resources

After you log in to WRDS website, you can use these resources:

- Main page: https://wrds-web.wharton.upenn.edu/wrds/
- WRDS Support Page
- 3 Ways to Use WRDS
- E-Learning Course Materials
- WRDS Research
- Sample Programs:

WRDS Sample Programs provide sample examples of data-extraction and cover only the basic aspects of data processing. Researchers can use these sample programs as a starting point to retrieve a sample of interest using WRDS data sources.

CRSP . Compustat (Legacy FTP) . OptionMetrics . TAQ . Thomson Reuters IBES . FirstCall . RiskMetrics . MFLinks, FISD . Trace . S&P Credit Ratings Global Insight . FDIC . CUSIP . CSMAR . Eventus . ComScore . Dow Jones . PHLX

6. Other Resources

- wrds.us : Wharton Research Data Services (WRDS) user community. This website contains tutorials, code samples, and forums geared towards using WRDS with SAS

- Codes & tools from Jie (Jay) Cao's website

- Data retrieval from CRSP and Compustat using Wharton's WRDS server from the Kellogg School of Management, Northwestern University

7. Data Sets

WRDS provides access to COMPUSTAT, CRSP, IBES, NYSE-TAQ, Bureau van Dijk, Global Insight, OptionMetrics, and other important business research databases.

A. COMPUSTAT

Compustat North America is a database of U.S. and Canadian fundamental and market information on active and inactive publicly held companies. It provides more than 300 annual and 100 quarterly Income Statement, Balance Sheet, Statement of Cash Flows, and supplemental data items. Compustat North America files are available in both annual and quarterly formats.

- Compustat Cross Reference (compare item names with legacy item numbers)
- Compustat Sample Programs

B. CRSP

The Center for Research in Security Prices (CRSP) maintains the most comprehensive collection of security price, return, and volume data for the NYSE, AMEX and NASDAQ stock markets. Additional CRSP files provide stock indices, beta- and cap-based portfolios, Treasury bond and risk-free rates, mutual funds, and real estate data.

- CRSP Sample Programs

C. Merging CRSP and Compustat:

You can look at these documents and codes:

- [PDF file] http://wrds-web.wharton.upenn.edu/wrds/E-Learning/_000Course%20Materials/Overview%20of%20CCM.pdf.cfm

- [code] wrds-web.wharton.upenn.edu/wrds/support/code_show.cfm?path=CRSP/CCM_XPF/merge_funda_crsp_byccm.sas

- [PDF file] wrds-web.wharton.upenn.edu/wrds/E-Learning/_000Course Materials/Merging_Compustat_CRSP.pdf.cfm

- [code] wrds-web.wharton.upenn.edu/wrds/E-Learning/_000Course Materials/CCMergeXPF.sas

8. EXAMPLES

COMPUSTAT

```
Program         :    COMPNA_FUNDA1.sas
Author          :    M Boldin /WRDS
Database        :    Compustat North America
```

```
/*****************************************
Remote Sign-on to WRDS Server
*****************************************/
%let wrds = wrds.wharton.upenn.edu 4016;
options comamid=TCP remote=WRDS;
signon username=_prompt_;

libname mylocal 'C:\Users\username\Desktop\SAS'; *define a local directory to
store output;

rsubmit;
options source nocenter ls=72 ps=max;
title 'Compustat North America data extract';
libname compx '/wrds/comp/sasdata/naa';

* Input Area: ************************************************************ ;
* Selected GVKEYS-- use quotes to be consistent with character variables;
%let glist = '006066' '012141'  '014489';

* Date range-- applied to FYEAR ( Fiscal Year);
%let fyear1= 1997;
%let fyear2= 2006;

*Selected data items (GVKEY, DATADATE, FYEAR and FYR are automatically
included);

%let vars=  cik tic at sale;
%let out_ds= work.compx2;

***  Start Program   ****************************************************** ;
*Make extract set of Compustat Annual company data;
data &out_ds;
   set compx.funda(keep= gvkey datadate fyear fyr &vars indfmt datafmt popsrc
consol);
   where gvkey in (&glist) and fyear between &fyear1 and &fyear2;
   if indfmt='INDL' and datafmt='STD' and popsrc='D' and consol='C';
   drop indfmt datafmt popsrc consol;  * Only used for the screening;
   sxa= sale/at;  * Sales over assets ratio;
run;
proc sort; by gvkey fyear; run;

proc print data=&syslast(obs=100);
  by gvkey;
  id fyear;
run;
```

MERGING CRSP AND COMPUSTAT

See:

- PowerPoint Slides: Merging CRSP and Compustat
- SAS Program Merge by Cusip

- SAS Program Merge using CCM table (See below)

STEPS

First, you need to remote sign on to WRDS (or alternatively you can work locally if you have already downloaded Compustat and CRSP files with link tables)

1. Extract Compustat data
2. Link GVKEYS to CRSP Identifiers using CCMXPF_LINKTABLE
3. Merge with CRSP data

```
/****************************************
Remote Sign on to WRDS Server
****************************************/
%let wrds = wrds.wharton.upenn.edu 4016;
options comamid=TCP remote=WRDS;
signon username=_prompt_;

libname mylocal 'C:\Desktop\'; *define a local directory to store output;

rsubmit;

* Define libraries for each database: CCM, CRSP monthly data, Compustat North
America;
libname crsp '/wrds/crsp/sasdata/cc';
libname crsp2 '/wrds/crsp/sasdata/sm';
libname comp '/wrds/comp/sasdata/na';


/*************************************************************
* STEP ONE: Extract Compustat data;
****************************************************************/

* Selected GVKEYS-- use quotes to be consistent with character variables;
%let glist = '006066' '012141'  '014489';

* Date range-- applied to FYEAR (Fiscal Year);
%let fyear1= 1997;
%let fyear2= 2006;

*  Selected data items (GVKEY, DATADATE, FYEAR and FYR are automatialy
included);
%let vars=  gvkey fyr fyear datadate SALE AT INDFMT DATAFMT POPSRC CONSOL;
```

```
* Make extract from Compustat Annual Funda file;
data compx2;
   set comp.funda (keep= &vars);
   where gvkey in (&glist) and fyear between &fyear1 and &fyear2;
   if indfmt='INDL' and datafmt='STD' and popsrc='D' and consol='C';
   * create begin and end dates for fiscal year;
   format endfyr begfyr date9.;
   endfyr= datadate;
   begfyr= intnx('month',endfyr,-11,'beg');  /* intnx(interval, from, n,
'aligment') */
   sxa= sale/at;  * compute sales over assets ratio;
   keep gvkey begfyr endfyr sxa fyr fyear;* keep only relevant variables;
run;

proc sort; by gvkey endfyr; run;

proc print data=&syslast(obs=100);
  by gvkey;
  id fyear;
  var fyear begfyr endfyr sxa;
run;

/**********************************************************************
* STEP TWO: Link GVKEYS to CRSP Identifiers;
* Use CCMXPF_LINKTABLE table to obtain CRSP identifiers for our subset of
companies/dates;
********************************************************************/

proc sql;
  create table mydata as select *
  from compx2 as a, crsp.ccmxpf_linktable as b
  where a.gvkey = b.gvkey and
  b.LINKTYPE in ("LU","LC","LD","LN","LS","LX") and
  b.usedflag=1 and
  (b.LINKDT <= a.endfyr or b.LINKDT = .B) and (a.endfyr <= b.LINKENDDT or
b.LINKENDDT = .E)
 /******************************************************************
  * The previous condition requires the end of fiscal year to fall within the
link range.       *
  *                                                                  *
  * A more relaxed condition would require any part of the fiscal year to be
within the link range:          *
  * (b.LINKDT <= a.endfyr or missing(b.LINKDT) = 1) and (b.LINKENDDT >=
a.begfyr or missing(b.LINKENDDT)= 1);*
  * or a more strict condition would require the entire fiscal year to be
within the link range :            *
  * (b.LINKDT <= a.begfyr or missing(b.LINKDT) = 1) and (a.endfyr <=
b.LINKENDDT or b.LINKENDDT= .E)          *
  *                                                                *
  * If these conditions are used, we suggest using the result data set from
the "collapsing" procedure -     *
  * which is shown in sample program ccm_lnktable.sas - to replace
crsp.ccmxpf_linktable.                   *
  *****************************************************************/
quit;
```

```
proc print data=&syslast(obs=50);
  by gvkey;
  id fyear;
run;

/*******************************************************
* STEP THREE: Add CRSP Monthly price data;
*************************************************************/
* Option 1: Simple match at the end of the fiscal year;

proc sql;
    create table mydata2 as select *
        from mydata as a, crsp2.msf as b
        where a.lpermno = b.permno and
        month(a.endfyr)=month(b.date) and year(a.endfyr)=year(b.date);
    quit;

proc print data=mydata2 (obs=30);
    var gvkey permno endfyr date sxa prc ret;
run;

/*********************************************************
* Option 2: Alternative way of matching CRSP data;
* Match accounting data with fiscal yearends in month 't',
  with CRSP return data from month 't+3' to month 't+14' (12 months);
*****************************************************/

proc sql;
create table mydata2 as select *
from mydata as a, crsp2.msf as b
where a.lpermno = b.permno and
      intck('month',a.endfyr,b.date) between 3 and 14;
quit;


proc download data=mydata2 out=mylocal.ccmfundaex; *download output dataset
to local location;
run;

endrsubmit;


proc print data=local.ccmfundaex (obs=30);
var gvkey permno endfyr date sxa prc ret;
run;
```

FAMA-FRENCH FACTORS

See: http://wrds-web.wharton.upenn.edu/wrds/research/applications/risk/fama-french/

Procedure (briefly)

1. Compustat

Use Compustat (annual data) as the source of historical accounting data in order to calculate Book Equity.

2. CRSP

Use CRSP monthly data. The first step in working with CRSP is to merge CRSP "event" and "time-series" files using a macro program (named 'crspmerge'). Second, add delisting returns (to reduce any bias in portfolio returns) and calculate Market Capitalization (ME) for each CRSP security (abs(prc)*shrout).

3. Merge CRSP and Compustat data

Merge CRSP and Compustat using the CRSP CCM product (linktable). Match Compustat's gvkey (from calendar year t-1) to CRSP's permno as of June year t.

After data cleaning, the book-to-market ratio for every firm in the sample were calculated by dividing Book Equity (for fiscal year that ends on year t-1) over the market value of its common equity at the end of December year t -1. These book-to-market ratios and Market Capitalization (as of December year t-1) were assigned to June year t in order to create portfolios.

4. Create Portfolios by Size and Book-to-Market. Calculation of FF factors

Every June (year t), calculate the median equity value of NYSE-listed firms using Market Capitalization at December year t-1. Use this median to classify firms as Small or Big in portfolios created at the end of June year t. In a similar fashion, as of June year t, firms are broken into three book-to-market equity groups (Low, Medium, and High) based on the 30% and 70% break-points of the NYSE-firms with positive book-to-market equity.

Subsequently, six size and book-to-market equity portfolios are created. Portfolios are created at the end of June and kept for 12 months. Within each portfolio, a monthly value-weighted return is calculated (in each month the weight is the Market Capitalization as of June year t adjusted by any change in price between the end June t and the end of the previous month).

The size factor, Small minus Big (SMB), is the difference between the average return on the three Small-firm portfolios and the average return on the three Big-firm portfolios. The value factor, High minus Low (HML), is the difference between the average return on the two High book-to-market equity portfolios and the average return on the two Low book-to-market equity portfolios.