

**COLLEGE OF MANAGEMENT**

**BUADM 722**

**CROSS-SECTIONAL ANALYSIS**

**FINANCIAL RESEARCH DATA**

1. **WRDS**



1. For registration for WRDS, <http://wrds-web.wharton.upenn.edu/>
2. WRDS allows you to access different databases including COMPUSTAT and CRSP.
3. For more detailed information regarding the variables
4. For sample SAS programming.
5. **COMPUSTAT DATA**

Compustat is a database of U.S. and Canadian fundamental and market information on more than 24,000 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.

**Annual Data:**

The example program below extracts data from Compustat for a group of firms, illustrates creation of lagged variables, and shows how new variables (ROA in this case) can be created.

1. These three lines establish a connection to the WRDS server where the data is stored. Substitute the user id and password for your personal WRDS account. The connection is terminated using a ***signoff*** statement.
2. ***Rsubmit*** and ***endrsubmit*** identify the beginning and end of the commands that will executed on the WRDS server. (libname comp "/wrds/comp/sasdata/naa";) is the physical address of the annual Compustat data in WRDS server.
3. The data set ‘comp.funda’ contains the Compustat annual data; the ***keep=*** list identifies which variables are to be retained from the full data set. A table of available variables are listed at the “support” section of the WRDS web site. Here *gvkey* (S&P permanent number), *cusip* (cusip), *tic* (stock ticker symbol), *fyear* (fiscal year), *fyr* (fiscal year end month), *at* (total assets), *sale* (sales) and *ib* (earnings before extraordinary items), *indfmt* (industry format), *datafmt* (data format), *porpsrc* (population source, *consol* (consolidation level) are selected, and *sich* (historical standard industrial classification code).

Note that the *sich* variable in Compustat is the SIC code in the observation year. In order to get the SIC code for the current period, we can take the *sich* of the most recent observation or we can find the sic variable in the dataset ‘comp.company’ (see next example).



1. ***IF*** or ***WHERE*** statements can be used to limit the sample to particular subsets of the data. Here 11 years of data are extracted, and financial firms (sic code 6xxx) are excluded.

Screening on *indfmt*='INDL', *datafmt*='STD', *popsrc*='D', and *consol*='C', will keep the vast majority of records and remove the duplicates from any *gvkey-fyear* pair. See Chapter 4 of **Compustat’s "Using the Data"** for detailed explanations of these additional codes.

1. Proc download transfers the data set to the local machine, where further processing is to be done. Once the data is transferred, the connection to WRDS is closed using ***signoff***.
2. These three lines create a lagged variable, *lag\_at*, which is the prior year-end assets. Notice that for the first year of this reduced dataset, the effect of this coding is to set the *lag\_at* variable to missing, which is denoted by a period (.) in SAS. Also note that care has been taken to avoid placing the lag operator in a conditional statement. The following code looks equivalent and more compact, but will NOT produce correct results:

IF cusip=lag(cusip) then lag\_at=lag(at);

1. ROA is earnings before extraordinary items (*ib*) divided by average assets. In general, when the division operator is used check first to make sure the denominator is not zero.
2. Saving the data in computer (libname zafer "C:\Users\zafer\Desktop";)
3. As a general practice, remember to print out a sample of observations to see if the results of the data extraction and manipulation are what you expected. Here the first 20 observations are printed, with the following result. Note that once the researcher has approved this result, the program can be modified to keep only the variables that will be used in subsequent processing.

**IMPORTANT NOTES: THE NAMES OF THE VARIABLES IN COMPUSTAT HAS BEEN CHANGED OVER TIME:** [**http://www.crsp.com/products/documentation/annual-data-industrial**](http://www.crsp.com/products/documentation/annual-data-industrial) **FOR CROSS REFERENCE.**



**Quarterly Data**

The next example program extracts data from quarterly Compustat for a group of firms, including earnings announcement date information.



1. The quarterly data is in ‘comp.fundq.’ The variable *rdq* is the earnings announcement date, *datadate* is the fiscal quarter end date, and *dlttq* is total long-term debt reported on the quarterly report.
2. In this example we are interested in companies with an SIC code between 3000 and 4000 in the year 2006. Since the ‘comp.fundq’ dataset does not have a *sich (SIC CODE)* variable, this *sich* variable must be attained from the ‘comp.funda’ dataset for the year of interest.
3. Here we use the Proc SQL command to merge the Comustat quarterly data with the SIC code data based on *gvkey*.



**Standard & Poors Execucomp**

The final database in this section is not technically Compustat, but is distributed by Standard and Poors, which also distributes Compustat. Execucomp includes compensation data collected from proxy disclosures of firms. The data includes cash compensation (salary and bonus), equity compensation (stock and option grants) and other information for each executive. Here is a little program to find the highest paid executives of 2006:



A. The main Execucomp data file is ‘anncomp.’ Here the retained variables include executive name, age and title, and compensation components, for 2006.

B. Cash compensation includes salary and bonus, and equity incentive compensation includes stock and option awards.

C. After sorting on total compensation, information about the ten highest paid executives. Total compensation for these fat cats ranged from $39.5 million to over $89 million, and four out of the top 10 were at Merrill Lynch:



**EXERCISES**

1. Use the Compustat annual dataset to select a sample of firms with 2006 R&D expense greater than 5% of sales. Use Proc Freq to tabulate the industries of the firms, and use Proc Univariate to look at the distribution of the R&D/Sales percentage.

2. Use the Compustat quarterly dataset to identify the 10% of 2006 firm-quarters with positive income before tax and lowest tax rates. Are these firms concentrated by industry? What is the mean, median, 25% decile, and 75% decile of the market value of these low tax rate firms?

3. Use the Execucomp data to identify the 10% of CEOs who got the largest percent increases in cash

1. **CRSP DATA**

The Center for Research in Security Prices, often known by its acronym CRSP, maintains security price, return, and volume data for the NYSE, AMEX and NASDAQ stock markets. Copies of the CRSP data manuals are available online in the Documentation and Manuals section of the WRDS web site. WRDS also provides numerous sample programs for executing common analysis task. In general, before you start programming, always check to see if one of the sample programs will help you.



The following program obtains market value (current and previous month), price (current and previous month), total share outstanding (current and previous month)



1. Variables needed for price, return and total share outstanding.
2. *cfacshr* and *cfacpr* are CRSP adjustment factors for share outstanding and stock price.
3. Creating Labels
4. Formatting the variables



1. **Merging Datasets**

One of the primary challenges in Accounting and Finance Research is the merging of data from different datasets. The major problem is that most of the datasets are organized based on proprietary primary identifiers that are not shared by other datasets. The identifiers used in different datasets are summarized in the following table.

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Primary Identifiers** | **Secondary Identifiers** |
| Compustat | GVKEY | Cusip, Ticker, Name, CIK |
| CRSP | PERMNO | Cusip, Ticker, Name |

Ideally, we would like to have a table that relates the Primary Indentifiers to each other; CRSP and Compustat collaborated together to construct such a table for *GVKEY* and *Permno*. This table is the basis of the Compustat/CRSP Merged (CCM) database. UMASS Boston currently subscribes to the CCM database, but not every university does. Therefore, the following sections will give examples of how to use the CCM database as well as how to create CRSP – Compustat link tables in the case that you do not have access to CCM.

In the absence of a primary key table, we must resort to use of one or more secondary identifiers, while recognizing that this approach may result in mismatched and unmatched data. The goal is to minimize these problems, and to the extent the problems are unavoidable, eliminate bad observations from the research data.



1. This ***Proc SQL*** is the basic code which links our Compustat data to a CRSP (*permno*) identifier, which can then be used to merge to CRSP data, in this case daily returns. Notice that in addition to simply linking by *gvkey*, we also make sure that the date of interest (*datadate*) falls within the date range for which the *gvkey* to *permno* link is effective. The designation “.B” in the *linkdt* field signifies that the link is effective since the beginning of the dates for which CRSP and Compustat overlap, and “.E” in the *linkenddt* means that the link is still effective as of the last update. We also restrict our linking to records in the CCM link table with a *usedflag*=1 and *linkprim*= ‘P’ or ‘C.’ These restrictions are necessary to avoid duplicate links or links that are inaccurate, per CCM documentation.