SAS for Data Management, Analysis, and Reporting
More on Proc Import and on File Merging

Other software packages

- Microsoft Excel
  - spreadsheet
  - very convenient for entering data in flat-file format
  - clients very frequently bring data to statisticians in Excel format
  - NOT reliable and accurate for doing statistical analysis

- Microsoft Access
  - relational database management system

Reading data files into SAS from other software packages

- Import Wizard
  - point-and-click interactive reading
  - convenient if file only needs to be read once
  - can write `proc import` code to be copied into programs

- `proc import`
  - can be used instead of data step in SAS programs
  - much more convenient if file needs to be read in multiple programs, or program using file needs to be run repeatedly

Importing from Other Sources

Types of files that the Import Wizard and/or `proc import` can read:

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Input Data Source</th>
<th>Extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCESS</td>
<td>Microsoft Access database</td>
<td>.MDB</td>
</tr>
<tr>
<td>DBF</td>
<td>dBASE file</td>
<td>.DBF</td>
</tr>
<tr>
<td>WK1</td>
<td>Lotus 1 spreadsheet</td>
<td>.WK1</td>
</tr>
<tr>
<td>WK3</td>
<td>Lotus 3 spreadsheet</td>
<td>.WK3</td>
</tr>
<tr>
<td>WK4</td>
<td>Lotus 4 spreadsheet</td>
<td>.WK4</td>
</tr>
<tr>
<td>EXCEL5</td>
<td>Excel V 5 spreadsheet</td>
<td>.XLS</td>
</tr>
<tr>
<td>EXCEL97</td>
<td>Excel 97 spreadsheet</td>
<td>.XLS</td>
</tr>
<tr>
<td>DLM</td>
<td>delimited file (default is blank)</td>
<td>.*</td>
</tr>
<tr>
<td>CSV</td>
<td>delimited file (comma-sep vals)</td>
<td>.CSV</td>
</tr>
<tr>
<td>TAB</td>
<td>delimited file (tab-delimited)</td>
<td>.TXT</td>
</tr>
</tbody>
</table>

Restriction: The data sources available to you depend on the SAS/ACCESS products that you have licensed. If you do not have any SAS/ACCESS products licensed, then the only types of data source files available to you are .CSV, .TXT, and delimited files.
Example

- from R or Splus
  - use `write.table` to write data out as a delimited file

Data frame that comes with R

```
> USArrests
         Murder Assault UrbanPop Rape
Alabama     13.2  236 58  21.2
Alaska      10.0  263 48  44.5
Arizona     8.1   294 80  31.0
Arkansas    8.8   190 50  19.5
California  9.0   276 91  40.6
```

R command to write out file as tab-delimited data file

```
> write.table(USArrests, file="C:\My Documents\166\USArrests.txt", sep="\t", quote = FALSE, col.names=TRUE)
```

Now in SAS....

File / Import Data

Import Wizard
Select a data source from the list below

Choose "Delimited File (*.*)"

Where is the file located?
Give full path name, e.g.
C:\My Documents\166\USArrests.txt

Choose SAS destination:
Library: (defaults to WORK)
Member: (fill in name of your choice; e.g. USArrest)

Question as to whether you want wizard to generate proc import statements so you can just run them next time

What it generated

```
PROC IMPORT OUT= WORK.usarrest DATAFILE= "C:\My Documents\166\USArrests.txt"
   DBMS=GLM REPLACE;
   DELIMITER='00'x; * needed correction to DELIMITED='09'x ;
   GETNAMES=YES;
   DATAROW=2;
RUN;
```

Example of reading Access database

```
PROC IMPORT OUT= WORK.courses DATATABLE= "Courses"
   DBMS=ACCESS97 REPLACE;
   DATABASE="c:\my documents\166\univ0_v7"
RUN;
```

More on file merging

Example: Acid rain deposition in Colorado

The National Atmospheric Deposition Program/National Trends Network (NADP/NTN) is a nationwide network of precipitation monitoring sites. The network is a cooperative effort between many different groups, including the State Agricultural Experiment Stations, U.S. Geological Survey, U.S. Department of Agriculture, and numerous other governmental and private entities. For a full list of contributors, see the collaborating agencies page. The NADP/NTN has grown from 22 stations at the end of 1978, our first year, to over 200 sites spanning the continental United States, Alaska, and Puerto Rico, and the Virgin Islands.

The purpose of the network is to collect data on the chemistry of precipitation for monitoring of geographical and temporal long-term trends. The precipitation at each station is collected weekly according to strict clean-handling procedures. It is then sent to the Central Analytical Laboratory where it is analyzed for hydrogen (acidity as pH), sulfate, nitrate, ammonium, chloride, and base cations (such as calcium, magnesium, potassium and sodium). Our excellent quality assurance programs ensure that the data remain accurate and precise.
We wish to report the following for all the NADP/NTN sites in Colorado:

1. site id code
2. elevation
3. sulfuric acid deposition in the year 2000 in kg/ha
4. actual number of days measurements contributing to year 2000 total
5. scatterplot of sulfuric acid deposition vs. elevation

We go to NADP/NTN data access at http://nadp.sws.uiuc.edu/

We download year 2000 annual data.

SAS code to read in this file

data depo ;
infile 'depoRepCO.asp' firstobs = 6 ;
input SiteID $QUOTE6. @9 Per $QUOTE8. Year Crit1 Crit2 Crit3 Crit4 Ca Mg K Na NH4 NO3 InorgN Cl SO4 HLab HField Svol Ppt Pct ValidF ValidL Days @201 Date1 $QUOTE12. @217 Date2 $QUOTE12. ;
drop Per Year Crit1-Crit4 ;
sdate = input(Date1, mmddyy10.) ; * convert from char var to date var ;
edate = input(Date2, mmddyy10.) ;
daysop = edate - sdate ;
format sdate edate date8. ;
run ;

Here is the output for the later variables in the dataset when format statement is not used.

<table>
<thead>
<tr>
<th>Obs</th>
<th>Date2</th>
<th>sdate</th>
<th>edate</th>
<th>daysop</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12/05/2000</td>
<td>12/28/1999</td>
<td>12/05/2000</td>
<td>343</td>
</tr>
<tr>
<td>2</td>
<td>01/02/2001</td>
<td>12/28/1999</td>
<td>01/02/2001</td>
<td>371</td>
</tr>
<tr>
<td>3</td>
<td>01/02/2001</td>
<td>12/28/1999</td>
<td>01/02/2001</td>
<td>371</td>
</tr>
<tr>
<td>4</td>
<td>01/02/2001</td>
<td>12/28/1999</td>
<td>01/02/2001</td>
<td>371</td>
</tr>
<tr>
<td>5</td>
<td>01/02/2001</td>
<td>12/29/1999</td>
<td>01/02/2001</td>
<td>370</td>
</tr>
</tbody>
</table>

Output when format statement is used.

Note:
- elevation is not in file
- header info in different format from actual records
Using “missover” and the “subsetting if” in data steps

data sites;
    infile 'stateCO.asp' firstobs = 19 missover;
    input SiteID $ 02 sitename $18. 040 strtdate mmddyy10. 053 stopdate ;
    if strtdate ne . ; * subsetting if: exclude observations meeting condition;
    format strtdate stopdate date8. ;
run ;

Combining the two datasets to produce the scatterplot

- Note that observations are in exactly the same order in the two datasets.
  - i.e. we want to match the first “site” record with the first “depo” record, the second with the second, etc.
- In this case only, we can use a one-to-one merge.

data combined;
    merge depo sites;
run;

proc print;
run;
Log file

data combined;
mmerge depo sites;
run;

NOTE: There were 17 observations read from the dataset WORK.DEP0.
NOTE: There were 18 observations read from the dataset WORK.SITES.
NOTE: The data set WORK.COMBINED has 18 observations and 27 variables.
NOTE: DATA statement used:
real time 0.08 seconds
cpu time 0.02 seconds

proc print;
run;

NOTE: There were 18 observations read from the dataset WORK.COMBINED.
NOTE: The PROCEDURE PRINT printed pages 1-2.
NOTE: PROCEDURE PRINT used:
real time 1.57 seconds
cpu time 0.07 seconds

Output from print

Site Inorg
Obs ID Ca Mg K Na NH4 NO3 N Cl SO4 HLab HField
1 C000 0.34 0.029 0.099 0.119 0.51 1.79 0.80 0.17 1.18 0.01 0.02
2 C001 1.07 0.076 0.103 0.138 1.36 3.09 1.76 0.19 2.06 0.00 0.01
3 C002 9.72 0.886 1.020 1.879 6.55 33.60 12.68 2.44 19.32 0.22 0.34
4 C008 0.97 0.098 0.128 0.152 0.57 3.93 1.33 0.26 2.13 0.03 0.05
5 C010 0.47 0.048 0.091 0.140 0.48 3.94 1.26 0.29 1.85 0.05 0.05
6 C015 0.73 0.081 0.036 0.150 0.45 3.35 1.11 0.22 1.80 0.03 -9.00
7 C019 0.81 0.076 0.121 0.145 0.73 3.79 1.72 0.21 2.00 0.02 0.03
8 C021 0.69 0.075 0.094 0.121 1.11 3.79 1.72 0.21 2.00 0.02 0.03
9 C022 0.74 0.075 0.120 0.138 0.57 3.93 1.33 0.26 2.13 0.03 0.05
10 C091 2.22 0.168 0.180 0.385 1.02 10.63 3.20 0.58 5.77 0.13 0.16
11 C092 1.16 0.118 0.111 0.221 0.75 5.55 1.84 0.34 2.99 0.06 0.06
12 C093 1.09 0.124 0.115 0.214 0.95 7.66 2.47 0.35 5.29 0.11 0.13
13 C094 0.74 0.074 0.109 0.145 0.45 3.35 1.11 0.22 1.80 0.03 9.00
14 C096 1.22 0.095 0.126 0.229 0.56 6.27 1.80 0.23 2.69 0.05 0.06
15 C097 1.58 0.163 0.136 0.313 1.63 11.55 3.88 0.52 8.62 0.18 0.20

Valid Valid

Obs Svol Ppt Pct F L Days Date1 Date2 sdate
1 6849.6 13.79 51 22 11 367 12/28/1999 12/05/2000 28DEC99
2 13377.7 23.85 83 27 22 371 12/28/1999 01/02/2001 28DEC99
3 26264.1 268.38 62 38 32 371 12/28/1999 01/02/2001 28DEC99
4 28189.3 49.05 70 43 29 371 12/28/1999 01/02/2001 28DEC99
5 26739.0 53.82 11 36 6 370 12/29/1999 01/02/2001 29DEC99
6 14207.8 32.56 0 32 0 371 12/28/1999 01/02/2001 28DEC99
7 20265.0 34.66 61 36 23 371 12/28/1999 01/02/2001 28DEC99
8 21956.6 34.24 79 30 22 371 12/28/1999 01/02/2001 28DEC99
10 40791.5 120.24 45 37 24 366 01/02/2000 01/02/2001 02JAN00
11 35244.8 69.20 63 42 27 371 12/28/1999 01/02/2001 28DEC99
12 25042.0 82.50 64 34 26 371 12/28/1999 01/02/2001 28DEC99
13 25603.0 41.33 97 40 35 371 12/28/1999 01/02/2001 28DEC99
14 33234.0 78.98 64 33 26 371 12/28/1999 01/02/2001 28DEC99
15 34487.0 135.99 71 36 26 371 12/28/1999 01/02/2001 28DEC99

Valid Valid

Obs edate daysop sitename strtdate stopdate elev
1 05DEC00 343 Alamosa 22APR80 . 2298
2 02JAN01 371 Las Animas Fish 04OCT83 . 1213
3 02JAN01 371 Hivot Saddle 05JUN84 . 3520
4 02JAN01 371 Four Mile Park 29DEC87 . 2502
5 02JAN01 370 Gothic 02FEB99 . 2926
6 02JAN01 371 Sand Spring 20MAR79 . 1998
7 02JAN01 371 Rocky Mtn National 29MAY80 . 2490
8 02JAN01 371 Manitou 17OCT78 . 2362
9 02JAN01 371 Pawnee 22MAR79 . 1641
10 02JAN01 366 Wolf Creek Pass 26MAY92 . 3292
11 02JAN01 371 Sunlight Peak 13JAN88 . 3206
12 02JAN01 371 Dry Lake 14OCT86 . 2527
13 02JAN01 371 Sugarloaf 04NOV86 . 2524
14 02JAN01 371 Nolan Pass 29JUL86 . 3249
15 02JAN01 371 Buffalo Pass 07FEB84 . 3234

Valid Valid

Obs Svol Ppt Pct F L Days Date1 Date2 sdate
16 28482.8 97.44 53 30 23 371 12/28/1999 01/02/2001 28DEC99
17 18208.4 30.54 52 30 18 371 12/28/1999 01/02/2001 28DEC99
18 . . . . . . . . . . . .
Looking at multiple records for each site

- Suppose we want to look at the annual sulfate ion deposition at the CO sites for each year from 1991-2000, inclusive
- We want to estimate site-specific random slopes on year, as well as fixed-effects intercept and coefficients of year and elevation
- Which SAS procedure?
- How should input data look?

- we need a “match merge”
- must process records in both files by a common variable
- then merge them by this variable
The log file

options linesize = 75 pagesize = 60 nodate nonumber;
data depo;
infile 'depoRep90s.asp' firstobs = 8;
input SiteID $ Per $8. Year Crit1 Crit2 Crit3 Crit4 Ca Mg
  K Na NH4 NO3 InorgN Cl SO4 HLab HField Svol Ppt Pct
  ! ValidF ValidL
.daysop = Date2 - Date1;
run;
NOTE: The infile 'depoRep90s.asp' is:
File
Name=/tmp_mnt/space/kcowles/166/lectures/lect2mkc/depoRep90s.asp,
Owner Name=kcowles,Group Name=faculty,
Access Permission=rw-------,
File Size (bytes)=35962
NOTE: 161 records were read from the infile 'depoRep90s.asp'.
The minimum record length was 218.
The maximum record length was 218.
NOTE: The data set WORK.DEPO has 161 observations and 7 variables.
NOTE: DATA statement used:
  real time 0.12 seconds
cpu time 0.07 seconds
Skipping stuff about sites file as we have seen it all.
data combined;
merge depo sites;
by SiteID;
run;
NOTE: There were 161 observations read from the dataset WORK.DEPO.
NOTE: There were 18 observations read from the dataset WORK.SITES.
NOTE: The data set WORK.COMBINED has 162 observations and 10 variables.
NOTE: DATA statement used:
  real time 0.02 seconds
cpu time 0.02 seconds

The SAS System

The combined file

<table>
<thead>
<tr>
<th>Site ID</th>
<th>Year</th>
<th>SO4</th>
<th>Daysop</th>
<th>Date1</th>
<th>Date2</th>
<th>daysop</th>
<th>strtdate</th>
<th>stopdate</th>
<th>elev</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO00</td>
<td>1991</td>
<td>2.08</td>
<td>364</td>
<td>01JAN91</td>
<td>31DEC91</td>
<td>364</td>
<td>22APR80</td>
<td>.</td>
<td>2298</td>
</tr>
<tr>
<td>CO00</td>
<td>1992</td>
<td>1.20</td>
<td>365</td>
<td>31DEC91</td>
<td>30DEC92</td>
<td>365</td>
<td>22APR80</td>
<td>.</td>
<td>2298</td>
</tr>
<tr>
<td>CO01</td>
<td>1991</td>
<td>3.19</td>
<td>363</td>
<td>02JAN91</td>
<td>31DEC91</td>
<td>363</td>
<td>04OCT83</td>
<td>.</td>
<td>1213</td>
</tr>
<tr>
<td>CO01</td>
<td>1992</td>
<td>3.09</td>
<td>364</td>
<td>31DEC91</td>
<td>29DEC92</td>
<td>364</td>
<td>04OCT83</td>
<td>.</td>
<td>1213</td>
</tr>
<tr>
<td>CO02</td>
<td>1991</td>
<td>9.88</td>
<td>365</td>
<td>29DEC98</td>
<td>28DEC99</td>
<td>364</td>
<td>29JUL86</td>
<td>.</td>
<td>2758</td>
</tr>
<tr>
<td>CO02</td>
<td>1992</td>
<td>9.46</td>
<td>365</td>
<td>29DEC98</td>
<td>28DEC99</td>
<td>364</td>
<td>29JUL86</td>
<td>.</td>
<td>3249</td>
</tr>
<tr>
<td>C094</td>
<td>1997</td>
<td>3.66</td>
<td>364</td>
<td>31DEC96</td>
<td>30DEC97</td>
<td>364</td>
<td>04NOV86</td>
<td>.</td>
<td>2524</td>
</tr>
<tr>
<td>C094</td>
<td>1998</td>
<td>4.34</td>
<td>364</td>
<td>31DEC96</td>
<td>30DEC97</td>
<td>364</td>
<td>04NOV86</td>
<td>.</td>
<td>2524</td>
</tr>
<tr>
<td>C094</td>
<td>1999</td>
<td>4.34</td>
<td>364</td>
<td>31DEC96</td>
<td>30DEC97</td>
<td>364</td>
<td>04NOV86</td>
<td>.</td>
<td>2524</td>
</tr>
<tr>
<td>C094</td>
<td>2000</td>
<td>2.79</td>
<td>371</td>
<td>31DEC96</td>
<td>02JAN91</td>
<td>371</td>
<td>04NOV86</td>
<td>.</td>
<td>2524</td>
</tr>
</tbody>
</table>
Omitting records missing from one file

data combined;
merge depo (in=ina) sites;
   * creates variable "ina" -- true if record is in depo, o.w. false;
by SiteID;
if ina;
   * subsetting if;
run;

Proc mixed

proc mixed data = combined;
class SiteID;
model so4 = year elev / s;
random year / subject = SiteID s;
run;

Proc transpose: exchanging rows and columns

Suppose instead we needed to process the data in the following format:
- a single row (record) for each site
- a column (variable) for each year’s so4 value

proc transpose data=combined out=combtran;
by SiteID;
id year;
var so4;
run;

proc print data=combtran (obs=10);
run;
More on “by” processing

- We can use certain SAS internal variables to extract certain observations within the “by groups.”

```sas
data firsts04 ;
set depo ;
by SiteID ;
if first.SiteID;
run ;

proc print data = firsts04 ;
run ;
```