22S:166
Computing in Statistics

More on merging

Lect 17
Nov. 1, 2006

Kate Cowles
374 SH, 335-0727
kcowles@stat.uiowa.edu

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


```
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 Days @196 Date1 mmddyy10. @209 Date2 mmddyy10. ;
drop Per Crit1-Crit4  Ca Mg K  Na  NH4  NO3 InorgN Cl HLab
HField  Svol  Ppt  Pct  ValidF ValidL ;
daysop = Date2 - Date1 ;
format Date2 Date1 date8. ;
run ;

*proc sort ;   * needed if records are not already in order ;
*by SiteID ;   * by SiteID ;
*run ;

data depo ;
set depo ;
by SiteID ;
run ;

proc print data  = depo (obs=25) ;
run ;

data sites ;
infile '/space/kcowles/166/lectures/lect1mkc/stateCO.asp' firstobs = 19
missover ;
input @13 SiteID $ @20 sitename $18. @40 strtdate mmddyy10.
@53 stopdate mmddyy10. @68 elev ;
if strtdate ne . ;   * subsetting if: exclude obs meeting condition ;
format strtdate stopdate date8. ;
drop sitename ;
```

```
run ;

* proc sort ;
* by SiteID ;
* run ;

data sites ;
set sites ;
by SiteID ;
run ;

* proc print ;
* run ;

data combined ;
merge depo sites ;
by SiteID ;
run ;

proc print data=combined ;
run ;
```


The log file

```
1          options linesize = 75 pagesize = 60 nodate nonumber ;
2
3          data depo ;
4          infile 'depoRep90s.asp' firstobs = 8 ;
5          input SiteID  $ Per $8. Year  Crit1 Crit2 Crit3 Crit4 Ca Mg
6          K  Na  NH4  NO3 InorgN Cl SO4 HLab  HField  Svol  Ppt  Pct
6        ! ValidF ValidL
7          Days @196 Date1 mmddyy10. @209 Date2 mmddyy10.  ;
8          drop Per Crit1-Crit4  Ca Mg K  Na  NH4  NO3 InorgN Cl HLab
8        ! HField
9          Svol  Ppt  Pct  ValidF ValidL ;
10         daysop = Date2 - Date1 ;
11         format Date2 Date1 date8. ;
12         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
```



```
13
14         *proc sort ;   * this is needed if records are not already in
14       ! order ;
15         *by SiteID ;   * by SiteID ;
16         *run ;
17
18         data depo ;
19         set depo ;
20         by SiteID ;
21         run ;

NOTE: There were 161 observations read from the dataset WORK.DEPO.
NOTE: The data set WORK.DEPO has 161 observations and 7 variables.
NOTE: DATA statement used:
      real time          0.01 seconds
      cpu time           0.02 seconds

Skipping stuff about sites file as we have seen it all.

48         data combined ;
49         merge depo sites ;
50         by SiteID ;
51         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

Obs	Site ID	Year	SO4	Days	Date1	Date2	daysop
1	CD00	1991	2.08	364	01JAN91	31DEC91	364
2	CD00	1992	1.20	365	31DEC91	30DEC92	365
3	CD00	1993	1.50	370	30DEC92	04JAN94	370
4	CD00	1994	1.31	364	04JAN94	03JAN95	364
5	CD00	1995	1.46	364	03JAN95	02JAN96	364
6	CD00	1996	1.07	364	02JAN96	31DEC96	364
7	CD00	1997	1.10	364	31DEC96	30DEC97	364
8	CD00	1998	1.28	364	30DEC97	29DEC98	364
9	CD00	1999	1.01	364	29DEC98	28DEC99	364
10	CD00	2000	1.18	367	28DEC99	05DEC00	343
11	CD01	1991	3.19	363	02JAN91	31DEC91	363
12	CD01	1992	3.09	364	31DEC91	29DEC92	364
13	CD01	1993	2.30	371	29DEC92	04JAN94	371
14	CD01	1994	2.98	364	04JAN94	03JAN95	364
15	CD01	1995	3.64	365	03JAN95	03JAN96	365
16	CD01	1996	2.99	363	03JAN96	31DEC96	363
17	CD01	1997	2.53	364	31DEC96	30DEC97	364
18	CD01	1998	2.44	364	30DEC97	29DEC98	364
19	CD01	1999	3.49	364	29DEC98	28DEC99	364
20	CD01	2000	2.06	371	28DEC99	02JAN01	371
21	CD02	1991	9.88	365	31DEC90	31DEC91	365
22	CD02	1992	9.46	364	31DEC91	29DEC92	364
23	CD02	1993	11.10	371	29DEC92	04JAN94	371
24	CD02	1994	7.96	364	04JAN94	03JAN95	364
25	CD02	1995	12.87	364	03JAN95	02JAN96	364

The combined file

Obs	Site ID	Year	SO4	Days	Date1	Date2	daysop	strtdate	stopdate	elev
1	CD00	1991	2.08	364	01JAN91	31DEC91	364	22APR80	.	2298
2	CD00	1992	1.20	365	31DEC91	30DEC92	365	22APR80	.	2298
3	CD00	1993	1.50	370	30DEC92	04JAN94	370	22APR80	.	2298
4	CD00	1994	1.31	364	04JAN94	03JAN95	364	22APR80	.	2298
5	CD00	1995	1.46	364	03JAN95	02JAN96	364	22APR80	.	2298
6	CD00	1996	1.07	364	02JAN96	31DEC96	364	22APR80	.	2298
7	CD00	1997	1.10	364	31DEC96	30DEC97	364	22APR80	.	2298
8	CD00	1998	1.28	364	30DEC97	29DEC98	364	22APR80	.	2298
9	CD00	1999	1.01	364	29DEC98	28DEC99	364	22APR80	.	2298
10	CD00	2000	1.18	367	28DEC99	05DEC00	343	22APR80	.	2298
11	CD01	1991	3.19	363	02JAN91	31DEC91	363	04OCT83	.	1213
12	CD01	1992	3.09	364	31DEC91	29DEC92	364	04OCT83	.	1213
13	CD01	1993	2.30	371	29DEC92	04JAN94	371	04OCT83	.	1213
14	CD01	1994	2.98	364	04JAN94	03JAN95	364	04OCT83	.	1213
15	CD01	1995	3.64	365	03JAN95	03JAN96	365	04OCT83	.	1213
16	CD01	1996	2.99	363	03JAN96	31DEC96	363	04OCT83	.	1213
17	CD01	1997	2.53	364	31DEC96	30DEC97	364	04OCT83	.	1213
18	CD01	1998	2.44	364	30DEC97	29DEC98	364	04OCT83	.	1213
19	CD01	1999	3.49	364	29DEC98	28DEC99	364	04OCT83	.	1213
20	CD01	2000	2.06	371	28DEC99	02JAN01	371	04OCT83	.	1213
.										
.										
.										
118	CD94	1997	3.66	364	31DEC96	30DEC97	364	04NOV86	.	2524
119	CD94	1998	4.34	364	30DEC97	29DEC98	364	04NOV86	.	2524
120	CD94	1999	3.61	364	29DEC98	28DEC99	364	04NOV86	.	2524
121	CD94	2000	2.79	371	28DEC99	02JAN01	371	04NOV86	.	2524
122	CD95	.	.	.	.	.	.	29JUL86	02JAN90	2758
123	CD96	1991	4.37	364	01JAN91	31DEC91	364	29JUL86	.	3249

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 ;
```