

22S:166 Computing in Statistics

More on merging

Sept. 17, 2001

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 mmdyy10. @209 Date2 mmdyy10. ;
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 mmdyy10.
@53 stopdate mmdyy10. @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 Sv1 Ppt Pct
7      ! ValidF ValidL
8      Days @196 Date1 mmdyyy10. @209 Date2 mmdyyy10. ;
9      drop Per Crit1-Crit4 Ca Mg K Na NH4 NO3 InorgN Cl HLab
10     ! HField
11     Sv1 Ppt Pct ValidF ValidL ;
12     daysop = Date2 - Date1 ;
13     format Date2 Date1 date8. ;
14     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	C000	1991	2.08	364	01JAN91	31DEC91	364
2	C000	1992	1.20	365	31DEC91	30DEC92	365
3	C000	1993	1.50	370	30DEC92	04JAN94	370
4	C000	1994	1.31	364	04JAN94	03JAN95	364
5	C000	1995	1.46	364	03JAN95	02JAN96	364
6	C000	1996	1.07	364	02JAN96	31DEC96	364
7	C000	1997	1.10	364	31DEC96	30DEC97	364
8	C000	1998	1.28	364	30DEC97	29DEC98	364
9	C000	1999	1.01	364	29DEC98	28DEC99	364
10	C000	2000	1.18	367	28DEC99	05DEC00	343
11	C001	1991	3.19	363	02JAN91	31DEC91	363
12	C001	1992	3.09	364	31DEC91	29DEC92	364
13	C001	1993	2.30	371	29DEC92	04JAN94	371
14	C001	1994	2.98	364	04JAN94	03JAN95	364
15	C001	1995	3.64	365	03JAN95	03JAN96	365
16	C001	1996	2.99	363	03JAN96	31DEC96	363
17	C001	1997	2.53	364	31DEC96	30DEC97	364
18	C001	1998	2.44	364	30DEC97	29DEC98	364
19	C001	1999	3.49	364	29DEC98	28DEC99	364
20	C001	2000	2.06	371	28DEC99	02JAN01	371
21	C002	1991	9.88	365	31DEC90	31DEC91	365
22	C002	1992	9.46	364	31DEC91	29DEC92	364
23	C002	1993	11.10	371	29DEC92	04JAN94	371
24	C002	1994	7.96	364	04JAN94	03JAN95	364
25	C002	1995	12.87	364	03JAN95	02JAN96	364

The combined file

```

Site
Obs  ID  Year  SO4  Days  Date1  Date2  daysop  strtdate  stopdate  elev
1  C000 1991  2.08  364  01JAN91 31DEC91  364  22APR80  . 2298
2  C000 1992  1.20  365  31DEC91 30DEC92  365  22APR80  . 2298
3  C000 1993  1.50  370  30DEC92 04JAN94  370  22APR80  . 2298
4  C000 1994  1.31  364  04JAN94 03JAN95  364  22APR80  . 2298
5  C000 1995  1.46  364  03JAN95 02JAN96  364  22APR80  . 2298
6  C000 1996  1.07  364  02JAN96 31DEC96  364  22APR80  . 2298
7  C000 1997  1.10  364  31DEC96 30DEC97  364  22APR80  . 2298
8  C000 1998  1.28  364  30DEC97 29DEC98  364  22APR80  . 2298
9  C000 1999  1.01  364  29DEC98 28DEC99  364  22APR80  . 2298
10 C000 2000  1.18  367  28DEC99 05DEC00  343  22APR80  . 2298
11 C001 1991  3.19  363  02JAN91 31DEC91  363  04OCT83  . 1213
12 C001 1992  3.09  364  31DEC91 29DEC92  364  04OCT83  . 1213
13 C001 1993  2.30  371  29DEC92 04JAN94  371  04OCT83  . 1213
14 C001 1994  2.98  364  04JAN94 03JAN95  364  04OCT83  . 1213
15 C001 1995  3.64  365  03JAN95 03JAN96  365  04OCT83  . 1213
16 C001 1996  2.99  363  03JAN96 31DEC96  363  04OCT83  . 1213
17 C001 1997  2.53  364  31DEC96 30DEC97  364  04OCT83  . 1213
18 C001 1998  2.44  364  30DEC97 29DEC98  364  04OCT83  . 1213
19 C001 1999  3.49  364  29DEC98 28DEC99  364  04OCT83  . 1213
20 C001 2000  2.06  371  28DEC99 02JAN01  371  04OCT83  . 1213
.
.
.
118 C094 1997  3.66  364  31DEC96 30DEC97  364  04NOV86  . 2524
119 C094 1998  4.34  364  30DEC97 29DEC98  364  04NOV86  . 2524
120 C094 1999  3.61  364  29DEC98 28DEC99  364  04NOV86  . 2524
121 C094 2000  2.79  371  28DEC99 02JAN01  371  04NOV86  . 2524
122 C095  .  .  .  .  .  .  .  .  02JAN90 2758
123 C096 1991  4.37  364  01JAN91 31DEC91  364  29JUL86  . 3249

```

11

Proc mixed

```

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

```

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 ;

```

12

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 ;

```

		Site										
Obs	ID	_NAME_	_1991_	_1992_	_1993_	_1994_	_1995_	_1996_	_1997_	_1998_	_1999_	_2000_
1	C000	S04	2.08	1.20	1.50	1.31	1.46	1.07	1.10	1.28	1.01	1.18
2	C001	S04	3.19	3.09	2.30	2.98	3.64	2.99	2.53	2.44	3.49	2.00
3	C002	S04	9.88	9.46	11.10	7.96	12.87	14.82	10.90	8.67	10.70	19.30
4	C008	S04	2.52	2.75	2.77	2.29	3.75	2.30	2.80	2.40	2.54	2.10
5	C010	S04	3.00	1.80
6	C015	S04	2.12	2.71	3.16	1.76	2.81	1.75	2.73	1.87	2.57	1.80
7	C019	S04	2.68	2.60	2.27	2.72	3.29	2.34	2.49	2.77	3.01	2.00
8	C021	S04	3.93	3.97	3.31	3.44	3.93	3.83	3.62	3.36	3.41	2.60
9	C022	S04	2.66	2.60	3.17	2.04	3.44	3.38	4.03	3.04	3.69	2.10
10	C091	S04	.	3.65	9.43	8.69	8.95	10.05	9.13	7.84	6.31	5.70

More on “by” processing

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

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

proc print data = firsts04 ;
run ;
```

		Site					
Obs	ID	Year	S04	Days	Date1	Date2	daysop
1	C000	1991	2.08	364	01JAN91	31DEC91	364
2	C001	1991	3.19	363	02JAN91	31DEC91	363
3	C002	1991	9.88	365	31DEC90	31DEC91	365
4	C008	1991	2.52	364	01JAN91	31DEC91	364
5	C010	1999	3.00	365	02FEB99	29DEC99	330
6	C015	1991	2.12	364	01JAN91	31DEC91	364
7	C019	1991	2.68	364	01JAN91	31DEC91	364
8	C021	1991	3.93	364	01JAN91	31DEC91	364
9	C022	1991	2.66	363	02JAN91	31DEC91	363
10	C091	1992	3.65	364	26MAY92	29DEC92	217
11	C092	1991	3.40	364	01JAN91	31DEC91	364
12	C093	1991	5.49	370	02JAN91	07JAN92	370
13	C094	1991	6.02	364	01JAN91	31DEC91	364
14	C096	1991	4.37	364	01JAN91	31DEC91	364
15	C097	1991	6.59	370	02JAN91	07JAN92	370
16	C098	1991	5.21	363	02JAN91	31DEC91	363
17	C099	1991	5.00	363	02JAN91	31DEC91	363