22S:166

Computing in Statistics
One approach to handling missing data
Arrays and looping in SAS

Lecture 28 Nov. 16, 2009

Kate Cowles
374 SH
kcowles@stat.uiowa.edu

Example dataset

ACTG 320 (Hammer, Squires, *et al.*, 1997) was a randomized, double-blind, placebo-controlled trial comparing a three-drug regimen (indinavir, lamivudine, and either zidovudine or stavudine) with a two-drug regimen (zidovudine and lamivudine) in HIV-infected adults with CD4 counts <= 200 and at least 3 months of prior zidovudine therapy. The 1156 randomized patients were stratified according to their CD4 count ($\leq$ 50 cells/mm^3 or 50-200 cells/mm^3) at study entry. The primary endpoint was occurrence of an AIDS-defining event (according to the CDC definition) or death. In addition, blood specimens were collected at baseline and at weeks 4, 8, 24, and 40 during follow-up for analysis of CD4 counts and viral load. The ACTG 320 dataset available for purchase from the National Technical Information Service includes clinical endpoints and CD4 data for all patients but viral load data on only 198 patients who were randomly selected for a virology substudy.

Example dataset, continued

- includes the 198 patients who have RNA data
- variables are

```
trt     -- 1/0 treatment group indicator
strat   -- 1/0 stratification group indicator
rna1    -- week 0 RNA
rna2    -- week 4 RNA
rna3    -- week 8 RNA
rna4    -- week 24 RNA
rna5    -- week 40 RNA
cd41    -- similar 5 cd4 values
cd42
cd43
cd44
cd45
obst    -- time at which clinical endpoint occurred, or last
           at which patient was observed and no clinical enc
fail    -- 1: cliical endpoint; 0: no clinical endpoint
```

- Note: no patient identifier

What we would like to do

- impute values for missing RNA and CD4 data
- calculate patient-specific rates of change of RNA by week and of CD4 by week
  - how will data file have to be laid out to do this?

Last-value-carried forward

- one (not terribly good) method of imputing missing values of longitudinal data
- may make sense if values are "missing at random"
  - that is, if the probability that a value is missing doesn't depend on the value that would have been observed
  - not likely to be the case for this type of data


```
options linesize = 72 ;

data actg320 ;
infile '/group/ftp/pub/kcowles/datasets/combo1.dat' firstobs = 2 ;
input trt strat rna1 rna2 rna3 rna4 rna5 cd41 cd42 cd43 cd44
      cd45 obst fail ;
pid = _N_ ;       * copy observation number into permanent variable ;
run ;

proc print data = actg320 (obs = 12) ;
title 'no arrays used' ;
run ;
```


no arrays used 1
18:26 Sunday, June 29, 2003

Obs	trt	strat	rna1	rna2	rna3	rna4	rna5
1	0	1	4.24790	3.27323	4.05660	3.98290	.
2	0	0	5.56951	5.10036	4.96781	5.41695	5.01041
3	0	0	4.96314	5.48520	4.60326	5.29003	5.96755
4	0	0	3.91666	3.53046	3.96881	2.03342	.
5	0	1	3.36286	2.29667	2.44560	3.55835	4.20850
6	1	1	4.65297	3.03342	2.71517	.	.
7	1	1	5.09670	2.87157	3.89856	2.14613	.
8	0	0	5.10824	5.14251	5.04452	2.37291	2.26007
9	1	0	5.52962	3.28825	2.88081	5.36319	.
10	1	1	4.97237	2.48001	2.13672	2.23045	2.30320
11	1	1	5.71936	3.45894	2.58995	2.25285	.
12	1	0	5.85708	3.18949	2.90309	2.02531	2.69984

Obs	cd41	cd42	cd43	cd44	cd45	obst	fail	pid
1	188.5	152	178	148	.	32.8571	0	1
2	19.0	43	35	7	.	43.8571	0	2
3	15.0	15	19	15	.	38.8571	0	3
4	30.0	30	40	70	.	49.7143	0	4
5	190.5	205	255	301	243	46.0000	0	5
6	128.5	166	867	.	.	24.8571	0	6
7	33.0	96	100	159	.	43.8571	0	7
8	10.0	10	11	72	67	13.0000	1	8
9	8.0	33	48	7	.	13.0000	1	9
10	139.0	178	119	305	305	48.7143	0	10
11	90.0	243	154	266	.	42.5714	0	11
12	20.0	155	172	160	142	49.0000	0	12

Arrays in SAS datasteps

- enable referencing a group of SAS variables by a single name and subscripts
- defined in `array` statements

```
array  arrayname[number of items] :
names or (list of
          values)
```

- exist during execution of data step

```
data actg320 ;
set actg320 ;
array arma[5] rna1 rna2 rna3 rna4 rna5 ;  * define array and make it a copy
                                            of existing variables in dataset
array acd4[5] cd41 cd42 cd43 cd44 cd45 ;
run ;

proc print data = actg320 (obs = 12) ;
title 'first way of doing arrays' ;
run ;
```

first way of doing arrays 2
18:26 Sunday, June 29, 2003

Obs	trt	strat	rna1	rna2	rna3	rna4	rna5
1	0	1	4.24790	3.27323	4.05660	3.98290	.
2	0	0	5.56951	5.10036	4.96781	5.41695	5.01041
3	0	0	4.96314	5.48520	4.60326	5.29003	5.96755
4	0	0	3.91666	3.53046	3.96881	2.03342	.
5	0	1	3.36286	2.29667	2.44560	3.55835	4.20850
6	1	1	4.65297	3.03342	2.71517	.	.
7	1	1	5.09670	2.87157	3.89856	2.14613	.
8	0	0	5.10824	5.14251	5.04452	2.37291	2.26007
9	1	0	5.52962	3.28825	2.88081	5.36319	.
10	1	1	4.97237	2.48001	2.13672	2.23045	2.30320
11	1	1	5.71936	3.45894	2.58995	2.25285	.
12	1	0	5.85708	3.18949	2.90309	2.02531	2.69984

Obs	cd41	cd42	cd43	cd44	cd45	obst	fail	pid
1	188.5	152	178	148	.	32.8571	0	1
2	19.0	43	35	7	.	43.8571	0	2

3	15.0	15	19	15	.	38.8571	0	3
4	30.0	30	40	70	.	49.7143	0	4
5	190.5	205	255	301	243	46.0000	0	5
6	128.5	166	867	.	.	24.8571	0	6
7	33.0	96	100	159	.	43.8571	0	7
8	10.0	10	11	72	67	13.0000	1	8
9	8.0	33	48	7	.	13.0000	1	9
10	139.0	178	119	305	305	48.7143	0	10
11	90.0	243	154	266	.	42.5714	0	11
12	20.0	155	172	160	142	49.0000	0	12


```
data actg320 ;
set actg320 ;
array arma[5] rna1 - rna5 ;
array acd4[5] cd41 - cd45 ;
run ;

proc print data = actg320 (obs = 12) ;
title 'second way of doing arrays' ;
run ;
```


second way of doing arrays 3
18:26 Sunday, June 29, 2003

Obs	trt	strat	rna1	rna2	rna3	rna4	rna5
1	0	1	4.24790	3.27323	4.05660	3.98290	.
2	0	0	5.56951	5.10036	4.96781	5.41695	5.01041
3	0	0	4.96314	5.48520	4.60326	5.29003	5.96755
4	0	0	3.91666	3.53046	3.96881	2.03342	.
5	0	1	3.36286	2.29667	2.44560	3.55835	4.20850
6	1	1	4.65297	3.03342	2.71517	.	.
7	1	1	5.09670	2.87157	3.89856	2.14613	.
8	0	0	5.10824	5.14251	5.04452	2.37291	2.26007
9	1	0	5.52962	3.28825	2.88081	5.36319	.
10	1	1	4.97237	2.48001	2.13672	2.23045	2.30320
11	1	1	5.71936	3.45894	2.58995	2.25285	.
12	1	0	5.85708	3.18949	2.90309	2.02531	2.69984

Obs	cd41	cd42	cd43	cd44	cd45	obst	fail	pid
1	188.5	152	178	148	.	32.8571	0	1
2	19.0	43	35	7	.	43.8571	0	2
3	15.0	15	19	15	.	38.8571	0	3
4	30.0	30	40	70	.	49.7143	0	4
5	190.5	205	255	301	243	46.0000	0	5
6	128.5	166	867	.	.	24.8571	0	6
7	33.0	96	100	159	.	43.8571	0	7
8	10.0	10	11	72	67	13.0000	1	8
9	8.0	33	48	7	.	13.0000	1	9
10	139.0	178	119	305	305	48.7143	0	10
11	90.0	243	154	266	.	42.5714	0	11
12	20.0	155	172	160	142	49.0000	0	12