

Comparing Values Across Observations: Be Careful

Prafulla Girase, IMS Health, Watertown, MA



Introduction

- Introduce three techniques that can be used to compare values across observations:
 - The LAG function
 - One-to-one reading
 - One-to-one merging
- Illustrate appropriate usage of each technique
- Discuss “gotchas” and advantages of the techniques

Sample Data

- The goal is to calculate time between visits for each patient

Output 1.1 Data Set VISITS

OBS	PAT_ID	VIS_NUM	VIS_DT
1	AA	1	01-Jan-03
2	AA	2	15-Feb-03
3	AA	3	18-Mar-03
4	BB	1	03-Apr-04
5	BB	2	15-Jun-04
6	BB	3	05-Jul-04
7	BB	4	10-Sep-04

1: The LAG function

- Stores a value in a queue and returns a value previously stored in a queue
- The intuitive definition is that it always returns the value of a variable from a previous observation (Gilson, SUGI 29)
- Storing values in a queue and returning values from a queue occurs only when the function is executed

1: The LAG function (cont.)

Output 1.1 Data Set VISITS

OBS	PAT_ID	VIS_NUM	VIS_DT	LAG_VIS_DT
1	AA	1	01-Jan-03	
2	AA	2	15-Feb-03	
3	AA	3	18-Mar-03	
4	BB	1	03-Apr-04	
5	BB	2	15-Jun-04	
6	BB	3	05-Jul-04	
7	BB	4	10-Sep-04	

QUEUE
.

1: The LAG function (cont.)

Output 1.1 Data Set VISITS

OBS	PAT_ID	VIS_NUM	VIS_DT	LAG_VIS_DT
1	AA	1	01-Jan-03	.
2	AA	2	15-Feb-03	
3	AA	3	18-Mar-03	
4	BB	1	03-Apr-04	
5	BB	2	15-Jun-04	
6	BB	3	05-Jul-04	
7	BB	4	10-Sep-04	

QUEUE
01-Jan-03

1: The LAG function (cont.)

Output 1.1 Data Set VISITS

OBS	PAT_ID	VIS_NUM	VIS_DT	LAG_VIS_DT
1	AA	1	01-Jan-03	.
2	AA	2	15-Feb-03	01-Jan-03
3	AA	3	18-Mar-03	
4	BB	1	03-Apr-04	
5	BB	2	15-Jun-04	
6	BB	3	05-Jul-04	
7	BB	4	10-Sep-04	

QUEUE
15-Feb-03

1: The LAG function (cont.)

Output 1.1 Data Set VISITS

OBS	PAT_ID	VIS_NUM	VIS_DT	LAG_VIS_DT
1	AA	1	01-Jan-03	.
2	AA	2	15-Feb-03	01-Jan-03
3	AA	3	18-Mar-03	15-Feb-03
4	BB	1	03-Apr-04	
5	BB	2	15-Jun-04	
6	BB	3	05-Jul-04	
7	BB	4	10-Sep-04	

QUEUE
18-Mar-03

1: The LAG function (cont.)

Output 1.1 Data Set VISITS

OBS	PAT_ID	VIS_NUM	VIS_DT	LAG_VIS_DT
1	AA	1	01-Jan-03	.
2	AA	2	15-Feb-03	01-Jan-03
3	AA	3	18-Mar-03	15-Feb-03
4	BB	1	03-Apr-04	18-Mar-03
5	BB	2	15-Jun-04	
6	BB	3	05-Jul-04	
7	BB	4	10-Sep-04	

QUEUE
03-Apr-04

1: The LAG function (cont.)

Output 1.1 Data Set VISITS

OBS	PAT_ID	VIS_NUM	VIS_DT	LAG_VIS_DT
1	AA	1	01-Jan-03	.
2	AA	2	15-Feb-03	01-Jan-03
3	AA	3	18-Mar-03	15-Feb-03
4	BB	1	03-Apr-04	18-Mar-03
5	BB	2	15-Jun-04	03-Apr-04
6	BB	3	05-Jul-04	
7	BB	4	10-Sep-04	

QUEUE
15-Jun-04

1: The LAG function (cont.)

Output 1.1 Data Set VISITS

OBS	PAT_ID	VIS_NUM	VIS_DT	LAG_VIS_DT
1	AA	1	01-Jan-03	.
2	AA	2	15-Feb-03	01-Jan-03
3	AA	3	18-Mar-03	15-Feb-03
4	BB	1	03-Apr-04	18-Mar-03
5	BB	2	15-Jun-04	03-Apr-04
6	BB	3	05-Jul-04	15-Jun-04
7	BB	4	10-Sep-04	

QUEUE
05-Jul-04

1: The LAG function (cont.)

Output 1.1 Data Set VISITS

OBS	PAT_ID	VIS_NUM	VIS_DT	LAG_VIS_DT
1	AA	1	01-Jan-03	.
2	AA	2	15-Feb-03	01-Jan-03
3	AA	3	18-Mar-03	15-Feb-03
4	BB	1	03-Apr-04	18-Mar-03
5	BB	2	15-Jun-04	03-Apr-04
6	BB	3	05-Jul-04	15-Jun-04
7	BB	4	10-Sep-04	05-Jul-04

QUEUE
10-Sep-04

1: The LAG function (cont.)

```
data incorrect ;  
  
set here.visits ;  
  
by pat_id vis_dt ;  
  
if not first.pat_id then LAG_VIS_DT =  
lag(vis_dt) ;  
  
label lag_vis_dt="Previous visit date" ;  
  
if not first.pat_id then diff = vis_dt -  
lag_vis_dt ;  
  
run ;
```

Output 1.2 Data Set INCORRECT

OBS	PAT_ID	VIS_DT	LAG_VIS_DT	DIFF
1	AA	01-Jan-03	.	.
2	AA	15-Feb-03	.	.
3	AA	18-Mar-03	15-Feb-03	31
4	BB	03-Apr-04	.	.
5	BB	15-Jun-04	18-Mar-03	455
6	BB	05-Jul-04	15-Jun-04	20
7	BB	10-Sep-04	05-Jul-04	67

1: The LAG function (cont.)

```
data correct ;  
  
  set here.visits ;  
  
  by pat_id vis_dt ;  
  
  LAG_VIS_DT = lag(vis_dt) ;  
  
  label lag_vis_dt="Previous visit date" ;  
  
  if first.pat_id then lag_vis_dt = . ;  
  
  if not first.pat_id then diff = vis_dt -  
    lag_vis_dt ;  
  
run ;
```

Output 1.3 Data Set CORRECT

OBS	PAT_ID	VIS_DT	LAG_VIS_DT	DIFF
1	AA	01-Jan-03	.	.
2	AA	15-Feb-03	01-Jan-03	45
3	AA	18-Mar-03	15-Feb-03	31
4	BB	03-Apr-04	.	.
5	BB	15-Jun-04	03-Apr-04	73
6	BB	05-Jul-04	15-Jun-04	20
7	BB	10-Sep-04	05-Jul-04	67

2: One-to-one reading

- Combines observations using two or more set statements
- The FIRSTOBS option can be used to create a lead function to compare values across observations
- The building of a new data set stops when any of the data sets being combined runs out of observations

2. One-to-one reading (cont.)

```
data incorrect ;
```

```
  set here.visits (firstobs=2
```

```
    rename=(vis_dt=NXT_VIS_DT))
```

```
  ;
```

```
  set here.visits ;
```

```
  by pat_id vis_dt ;
```

```
  if last.pat_id then nxt_vis_dt= . ;
```

```
  if not last.pat_id then diff = nxt_vis_dt -  
    vis_dt ;
```

```
run ;
```

Output 1.4 Data Set Incorrect

OBS	PAT_ID	VIS_DT	NXT_VIS_DT	DIFF
1	AA	01-Jan-03	15-Feb-03	45
2	AA	15-Feb-03	18-Mar-03	31
3	AA	18-Mar-03	.	.
4	BB	03-Apr-04	15-Jun-04	73
5	BB	15-Jun-04	05-Jul-04	20
6	BB	05-Jul-04	10-Sep-04	67

2. One-to-one reading (cont.)

```
data correct ;
```

```
if not eof then do ;
```

```
set here.visits (firstobs=2
```

```
rename=(vis_dt=NXT_VIS_DT))  
end=eof ;
```

```
end ;
```

```
set here.visits ;
```

```
by pat_id vis_dt ;
```

```
if last.pat_id then nxt_vis_dt= . ;
```

```
if not last.pat_id then diff = nxt_vis_dt -  
vis_dt ;
```

```
run ;
```

Output 1.5 Data Set Correct

OBS	PAT_ID	VIS_DT	NXT_VIS_DT	DIFF
1	AA	01-Jan-03	15-Feb-03	45
2	AA	15-Feb-03	18-Mar-03	31
3	AA	18-Mar-03	.	.
4	BB	03-Apr-04	15-Jun-04	73
5	BB	15-Jun-04	05-Jul-04	20
6	BB	05-Jul-04	10-Sep-04	67
7	BB	10-Sep-04	.	.

3. One-to-one merging

- A MERGE without a BY statement
- A classic way of doing the lead (Dunn, PharmaSUG TU09)
- The number of observations in the final data set is equal to the number of observations in the largest data set being combined
- The FIRSTOBS option can be used to create a lead function to compare values across observations

3. One-to-one merging (cont.)

```
data correct ;  
  
merge here.visits (firstobs=2 keep=vis_dt  
                  rename=(vis_dt=NXT_VIS_DT))  
      here.visits ;  
  
run ;  
  
data correct ;  
  
set correct ;  
  
by pat_id vis_dt ;  
  
if last.pat_id then nxt_vis_dt= . ;  
  
if not last.pat_id then diff = nxt_vis_dt -  
    vis_dt ;  
  
run ;
```

Output 1.6 Data Set CORRECT

OBS	PAT_ID	VIS_DT	NXT_VIS_DT	DIFF
1	AA	1-Jan-03	15-Feb-03	45
2	AA	15-Feb-03	18-Mar-03	31
3	AA	18-Mar-03	.	.
4	BB	3-Apr-04	15-Jun-04	73
5	BB	15-Jun-04	5-Jul-04	20
6	BB	5-Jul-04	10-Sep-04	67
7	BB	10-Sep-04	.	.

Conclusion

- No one technique is right or wrong
- A good understanding of your data and how each technique works is important when selecting an appropriate technique
- One-to-one reading and merging both provide lead values while the first technique provides lag values
- Between one-to-one reading and merging, merging is recommended since it is easy to implement and less prone to errors

My Contact info

- Prafulla Girase
- Email: pgirase@us.imshealth.com
- Phone: 617-393-8357

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks