

Supplement: The IN= Statement

This supplement is offered in response to a question a few students raised with regard to the IN= statement. Please refer to the course notes/textbook for more details on this statement and its purpose.

Some of you were asking about the example we had in class where we did NOT use two (IN=...) statements ... although in the in-class activity, you'll find that in some instances you DO need two (IN=...) statements.

The answer is: it never hurts to have two (IN=...) statements when merging two data sets. However, when **ONE DATA SET IS COMPLETELY CONTAINED IN THE OTHER**, then you really don't need to have an (IN=...) statement attached to the larger set. Let me explain.

Example: A company keeps track of certain periods of the calendar year. The company goes through a review process every other month starting in January. Also, the busiest months for this company are from September through January.

Suppose we have the following data sets called CAL (for calendar), REV (for review), and BUSY. I have intentionally listed the data sets with gaps so that we can easily see where there is overlap between the sets and where there isn't overlap.

CAL	REV	BUSY																																																										
<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left;">ID</th> <th style="text-align: left;">MON</th> </tr> </thead> <tbody> <tr> <td>==</td> <td>===</td> </tr> <tr> <td>01</td> <td>Jan</td> </tr> <tr> <td>02</td> <td>Feb</td> </tr> <tr> <td>03</td> <td>Mar</td> </tr> <tr> <td>04</td> <td>Apr</td> </tr> <tr> <td>05</td> <td>May</td> </tr> <tr> <td>06</td> <td>Jun</td> </tr> <tr> <td>07</td> <td>Jul</td> </tr> <tr> <td>08</td> <td>Aug</td> </tr> <tr> <td>09</td> <td>Sep</td> </tr> <tr> <td>10</td> <td>Oct</td> </tr> <tr> <td>11</td> <td>Nov</td> </tr> <tr> <td>12</td> <td>Dec</td> </tr> </tbody> </table>	ID	MON	==	===	01	Jan	02	Feb	03	Mar	04	Apr	05	May	06	Jun	07	Jul	08	Aug	09	Sep	10	Oct	11	Nov	12	Dec	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left;">ID</th> <th style="text-align: left;">MON</th> </tr> </thead> <tbody> <tr> <td>==</td> <td>===</td> </tr> <tr> <td>01</td> <td>Jan</td> </tr> <tr> <td>03</td> <td>Mar</td> </tr> <tr> <td>05</td> <td>May</td> </tr> <tr> <td>07</td> <td>Jul</td> </tr> <tr> <td>09</td> <td>Sep</td> </tr> <tr> <td>11</td> <td>Nov</td> </tr> </tbody> </table>	ID	MON	==	===	01	Jan	03	Mar	05	May	07	Jul	09	Sep	11	Nov	<table style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left;">ID</th> <th style="text-align: left;">MON</th> </tr> </thead> <tbody> <tr> <td>==</td> <td>===</td> </tr> <tr> <td>01</td> <td>Jan</td> </tr> <tr> <td>09</td> <td>Sep</td> </tr> <tr> <td>10</td> <td>Oct</td> </tr> <tr> <td>11</td> <td>Nov</td> </tr> <tr> <td>12</td> <td>Dec</td> </tr> </tbody> </table>	ID	MON	==	===	01	Jan	09	Sep	10	Oct	11	Nov	12	Dec
ID	MON																																																											
==	===																																																											
01	Jan																																																											
02	Feb																																																											
03	Mar																																																											
04	Apr																																																											
05	May																																																											
06	Jun																																																											
07	Jul																																																											
08	Aug																																																											
09	Sep																																																											
10	Oct																																																											
11	Nov																																																											
12	Dec																																																											
ID	MON																																																											
==	===																																																											
01	Jan																																																											
03	Mar																																																											
05	May																																																											
07	Jul																																																											
09	Sep																																																											
11	Nov																																																											
ID	MON																																																											
==	===																																																											
01	Jan																																																											
09	Sep																																																											
10	Oct																																																											
11	Nov																																																											
12	Dec																																																											

(1) Find the list of non-review months:

```

SAS Code
data nonreview;
  merge CAL REV (in = INrev);
  by ID;
  if INrev=0 then output;
SAS Code

```

When we MERGE the data sets CAL and REV, notice that we simply obtain CAL once again since the smaller set REV is **completely contained** in the CAL set.

CAL		REV		CAL&REV MERGED		
ID	MON	ID	MON	ID	MON	INrev
==	===	==	===	==	===	=====
01	Jan	01	Jan	01	Jan	1
02	Feb			02	Feb	0
03	Mar	03	Mar	03	Mar	1
04	Apr			04	Apr	0
05	May	05	May	05	May	1
06	Jun			06	Jun	0
07	Jul	07	Jul	07	Jul	1
08	Aug			08	Aug	0
09	Sep	09	Sep	09	Sep	1
10	Oct			10	Oct	0
11	Nov	11	Nov	11	Nov	1
12	Dec			12	Dec	0

So, from the perspective of the merged data set (found above and to the right), we can simply pick out the **non-review** months by selecting all observations with INrev=0.

You could have used

```

SAS Code
data nonreview;
  merge CAL (in=INcal) REV (in = INrev);
  by ID;
  if INcal=1 and INrev=0 then output;
SAS Code

```

but that would be redundant since CAL and the merged set are identical! So, the value of INcal would be 1 for all observations in the merged set. Meaning, the first condition of the if statement (INcal=1) would *always* be satisfied ... so it's really not necessary.

(2) Find the list of review months that are busy:

Notice that REV is not totally contained in BUSY (and vice versa). REV contains observations not found in BUSY.

REV		BUSY		REV&BUSY MERGED			
ID	MON	ID	MON	ID	MON	INrev	INbusy
==	===	==	===	==	===	=====	=====
01	Jan	01	Jan	01	Jan	1	1
03	Mar			03	Mar	1	0
05	May			05	May	1	0
07	Jul			07	Jul	1	0
09	Sep	09	Sep	09	Sep	1	1
		10	Oct	10	Oct	0	1
11	Nov	11	Nov	11	Nov	1	1
		12	Dec	12	Dec	0	1

When we MERGE the data sets REV and BUSY, notice that we cannot just rely upon **one** of the indicator variables. For example, suppose we used:

```

SAS Code
data revbusy1;
  merge REV BUSY (in = INbusy);
  by ID;
  if INbusy=1 then output;
SAS Code

```

It should be obvious from the merged data set (shown above and to the right) that this code would yield **ONLY** the list of busy months! This is not what we want.

```

SAS Code
data revbusy2;
  merge REV (in = INrev) BUSY;
  by ID;
  if INrev=1 then output;
SAS Code

```

Here, it should be obvious that this code would yield **ONLY** the list of review months! This is not what we want.

SAS Code

```
data revbusy3;
  merge REV (in = INrev) BUSY (in = INbusy);
  by ID;
  if INrev=1 and INbusy=1 then output;
```

SAS Code

Ahhh ... now that seems to be better! We want **ONLY** those observations that are review and busy months. We need **BOTH** indicator variables.