Chapter 5

Combining datasets in SAS

In this chapter we will look at some different ways to combine datasets. As examples, we shall return to the analysis of the two diabetic datasets 'drugE.txt' and 'drugN.txt'. Remember the format of each dataset is

- number  patient id
- ssbp_0  baseline supine systolic blood pressure
- ssbp_14 supine systolic blood pressure after 14 days
- sdbp_0  baseline supine diastolic blood pressure
- sdbp_14 supine diastolic blood pressure after 14 days

Make sure both datasets are stored in the folder My Computer → Drive C: → My SAS files.

5.1 Modifying datasets and stacking datasets

We can create new datasets using the DATA statement with the SET command:

DATA new_data;
SET old_data;

This command creates a new dataset new_data from the old dataset old_data. We can then create variables in this new dataset as normal. The SET command is also useful when we want to stack two datasets on top of each other. Here is a program to stack the drugE dataset on top of the drugN dataset.
/* Load the 'drugE' dataset */
DATA drugE;
   INFILE "C:\My SAS Files\drugE.txt" FIRSTOBS=2;
   INPUT number ssbp_0 ssbp_14 sdbp_0 sdbp_14;
   drug = "E";

/* Load the 'drugN' dataset */
DATA drugN;
   INFILE "C:\My SAS Files\drugN.txt" FIRSTOBS=2;
   INPUT number ssbp_0 ssbp_14 sdbp_0 sdbp_14;
   drug = "N";

/* stack drugE on top of drugN */
DATA diabetics;
   SET drugE drugN;

PROC PRINT data=diabetics NOOBS;
RUN;

Exercise: Write a program which DROPs the systolic blood pressure at day 0 variable from the 'drugE' dataset only. Now stack this new 'drugE' dataset on top of the dataset 'drugN'. What is the effect of missing out a variable? What does the resulting stacked dataset look like?

5.2 Interleaving datasets

We can add the BY command to the last DATA statement of the previous to interleave the two datasets by some variable. By interleave, we mean stacking the two datasets on top of each other, and then sorting this new dataset by some variable in the stacked dataset. To use the BY command, the two datasets need to be already sorted by the variable. Here is a program interleaves the two datasets 'drugE' and 'drugN' by the patient number.

/* Load the 'drugE' dataset */
DATA drugE;
   INFILE "C:\My SAS Files\drugE.txt" FIRSTOBS=2;
   INPUT number ssbp_0 ssbp_14 sdbp_0 sdbp_14;
   drug = "E";

/* sort 'drugE' by number - call the resulting dataset 'drugE_by_number' */
PROC SORT DATA=drugE OUT=drugE_by_number;
   BY number;
/* Load the 'drugN' dataset */
DATA drugN;
   INFILE "C:\My SAS Files\drugN.txt" FIRSTOBS=2;
   INPUT number ssbp_0 ssbp_14 sdbp_0 sdbp_14;
   drug = "N";

/* sort 'drugN' by number - call the resulting dataset 'drugN_by_number' */
PROC SORT DATA=drugN OUT=drugN_by_number;
   BY number;

/* stack 'drugE_by_number' on top of 'drugN_by_number' */
DATA diabetics;
   SET drugE_by_number drugN_by_number;
   BY number;

PROC PRINT data=diabetics NOOBS;
RUN;

Exercise: Write a program to interleaved the two datasets, sorted by the systolic blood pressure measured at day 0.

5.3 Merging datasets

Suppose we have age data for each patient number in the 'drugE' dataset. Here is a program to define the age data.

DATA drugE_age;
   INPUT number age;
   DATALINES;
21 23
8 36
19 31
20 17
24 25
9 18
12 29
16 32
;  
RUN;
We use the **MERGE** command in the **DATA** statement when we wish to combine two datasets which have a common variable. The format of the command is

```
DATA dataset;
   MERGE to_merge1 to_merge2;
   BY variables;
```

The command will merge the data from `to_merge1` and `to_merge2` using the variables `variables` and store the result in `dataset`. The datasets `to_merge1` and `to_merge2` have to be already sorted by `variables`. Here is the program to merge the age data with the original `drugE` dataset.

```sas
/* Load the 'drugE' dataset */
DATA drugE;
   INFILE "C:\My SAS Files\drugE.txt" FIRSTOBS=2;
   INPUT number ssbp_0 ssbp_14 sdbp_0 sdbp_14;
   drug = "E";

/* sort 'drugE' by number - call the resulting dataset 'drugE_by_number' */
PROC SORT DATA=drugE OUT=drugE_by_number;
   BY number;

/* create the age dataset. */
DATA drugE_age;
   INPUT number age;
   DATALINES;
21 23
8 36
19 31
20 17
24 25
9 18
12 29
16 32
;```

```sas
/* sort the age dataset by number */
PROC SORT DATA=drugE_age OUT=drugE_age_by_number;
   BY number;
```
/ * merge the two datasets */
DATA drugE_merged;
    MERGE drugE_by_number drugE_age_by_number;
    BY number;

PROC PRINT DATA=drugE_merged NOOBS;
RUN;

**Exercise 1**: Change the DATALINES above to read

```
DATALINES;
21 23
8 36
19 31
;
```

What does the merged dataset look like?

**Exercise 2**: Change part of the above program to read

```
/* create the age dataset. */
DATA drugE_age;
    INPUT number age ssbp_0;
    DATALINES;
21 23 2
8 36 3
19 31 4
20 17 5
24 25 6
9 18 7
12 29 8
16 32 9
;
```

What happens when we merge two datasets that have the same variable names, but have different contents? (note here that ssbp_0 is not a BY variable in this program). Try changing the order of the MERGE command, i.e., change the MERGE to

```
MERGE drugE_age_by_number drugE_by_number;
```