Chapter 12 READING SAS DATA SETS

FINDING FIRST AND LAST OBSERVATIONS IN SUBGROUPS

* **FIRST**.*variable* is set to 1 at the first occurrence of a new value for a sorted *variable*; otherwise, it is set to 0.
* **LAST**.*variable* is set to 1 at the last occurrence of a new value for a sorted *variable*; otherwise, it is set to 0.
* Note that both FIRST.variable and LAST.variable are temporary variables that are not stored in the new data set.

Example. In the exercise example, the following program calculates total payments due for men and for women.

proc sort data=exercise; /\* data set has to be sorted \*/

by gender;

run;

data exercise;

set exercise;

by gender; /\* defines subgroups \*/

if first.gender=1 then total\_payment=0;

total\_payment+payment;

if last.gender=1; /\* selects only last observation in each gender category \*/

run;

proc print noobs;

run;

The output is

total\_

id age actlevel gender payment payment

2833 32 MOD F 15.5 78.50

2823 29 HIGH M 23.0 70.75

Underneath the surface SAS computes the following:



* Only the last observation in each gender category is output.
* The variables First.gender and Last.gender are not printed.

READING OBSERVATIONS USING DIRECT ACCESS

The **POINT=** option in the SET statement may be used to access observations directly by their observation number. The syntax is

POINT = *variable\_name*;

where *variable\_name* specifies a temporary numeric variable that contains the observation number of the observation to be read. It must be given a value before the SET statement is executed.

Example. In the exercise example, only the third observation is read by the following code.

data observation3;

obsnum=3;

set exercise point=obsnum;

output;

stop;

run;

proc print data=observation3 noobs;

run;

The output is

id age actlevel gender payment

2807 42 LOW M 27.5

* The OUTPUT statement stores the read observation in the new data set. If it is not specified, the new data set will be empty.
* The STOP statement terminates the sequential data reading by the SET statement. If it is not specified, SAS goes into infinite looping.

DETECTING THE END OF A DATA SET

The **END=** option in the SET statement may be used to detect when the last observation in an input data set is read. The syntax is

END = *variable\_name*;

where *variable\_name* is set to 1 if when SAS reads the last observation of a data set, and is set to 0 otherwise. It is a temporary variable not stored in the new data set.

* The END= option is useful when the total number of observations in a data set is **unknown**.
* If the total number of observations in a data set is **known**, the POINT= option may be used as an alternative.

Example. To output the last observation in the exercise data set, the following program may be run.

data exercise;

set exercise end=last; /\*last is an end-of-file marker\*/

if last=1; /\* selects only the last observation \*/

run;

proc print noobs;

run;

The output is

id age actlevel gender payment

2823 29 HIGH M 23

Chapter 20 CREATING A SINGLE OBSERVATION

FROM MULTIPLE RECORDS

Observations for one individual may be spread out over a number of lines. There are several ways to read each record.

Example. The following data contain individual’s name, number and street name, city, state, and ZIP.

Nguyen, Melany 6249 ELM AVENUE CYPRESS CA 90630

Mari, Leana Adrienne 3199 MAIN STREET SANTA MONICA CA 90405

Cawley, Anne Margaret 277 N VERMONT AVENUE LOS ANGELES CA 90004

* The data may be read using **multiple INPUT statement**.

data address;

input name $21.;

input number $4. @6 street $20.;

input city $12. @14 state $2. @17 zip $5.;

cards;

Nguyen, Melany

6249 ELM AVENUE

CYPRESS CA 90630

Mari, Leana Adrienne

3199 MAIN STREET

SANTA MONICA CA 90405

Cawley, Anne Margaret

277 N VERMONT AVENUE

LOS ANGELES CA 90004

;

proc print noobs;

run;

The output is

name number street city state zip

Nguyen, Melany 6249 ELM AVENUE CYPRESS CA 90630

Mari, Leana Adrienne 3199 MAIN STREET SANTA MONICA CA 90405

Cawley, Anne Margaret 277 N VERMONT AVENUE LOS ANGELES CA 90004

* The data may be read using a **single** **INPUT statement** containing a **line pointer control (#, sharp)** that specifies the number of the line from which values are to be read.

data address;

input #1 name $21.

#2 number $4. @6 street $20.

#3 city $12. @14 state $2. @17 zip $5.;

cards;

. . .

;

* The data may be read using a **single INPUT statement** containing the **forward slash (/) line pointer control** that tells SAS to advance to the next line.

data address;

input name $21./

number $4. @6 street $20./

city $12. @14 state $2. @17 zip $5.;

cards;

. . .

;

* The data may be read **non-sequentially** using a **single** **INPUT statement** containing a **line pointer control (#).**

data address;

input #2 number $4. @6 street $20.

#3 city $12. @14 state $2. @17 zip $5.

#1 name $21.;

cards;

. . .

;

proc print noobs;

run;

The output is

number street city state zip name

6249 ELM AVENUE CYPRESS CA 90630 Nguyen, Melany

3199 MAIN STREET SANTA MONICA CA 90405 Mari, Leana Adrienne

277 N VERMONT AVENUE LOS ANGELES CA 90004 Cawley, Anne Margaret