Summary of Lesson 1: Controlling DATA Step Processing

Understanding DATA Step Processing

- The DATA step is processed in two phases: compilation and execution.
- During compilation, SAS creates the program data vector (PDV) and establishes data attributes and rules for execution.
- The PDV is an area of memory established in the compilation phase. It includes all columns that will be read or created, along with their assigned attributes. The PDV is used in the execution phase to hold and manipulate one row of data at a time.
- During execution, SAS reads, manipulates, and writes data. All data manipulation is performed in the PDV.

PUTLOG _ALL_;
PUTLOG *column=;*
PUTLOG *"message";*

Directing DATA Step Output

OUTPUT;
DATA *table1 <table2 ...>*;
OUTPUT *table1 <table2 ...>*;

- By default, the end of a DATA step causes an implicit OUTPUT, which writes the contents of the PDV to the output table.
- The explicit OUTPUT statement can be used in the DATA step to control when and where each row is written.
- If an explicit OUTPUT statement is used in the DATA step, it disables the implicit OUTPUT at the end of the DATA step.
- One DATA step can create multiple tables by listing each table name in the DATA statement.
- The OUTPUT statement followed by a table name writes the contents of the PDV to the specified table.

table **(DROP=***col1 col2…***);**
table **(KEEP=***col1 col2…***);**

- DROP= or KEEP= data set options can be added on any table in the DATA statement. If you add the DROP= option, the columns that you list are not added to the output table. If you add the KEEP= option, **only** the columns that you list are added to the output table.

- Columns that will be dropped are flagged in the PDV and are not dropped until the row is output to the designated table. Therefore, dropped columns are still available for processing in the DATA step.

- DROP= or KEEP= data set options can be added in the SET statement to control the columns that are read into the PDV. If a column is not read into the PDV, it is not available for processing in the DATA step.