



Don't Be a SAS® Dinosaur: Modernize Your SAS Programs

by Warren Repole

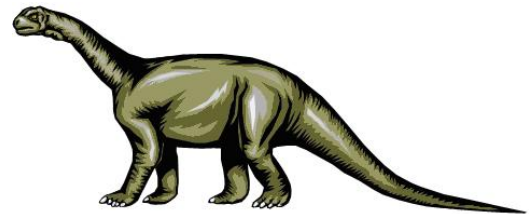
Duplicate Observations: Remove Complete Duplicates

Scenario:

You need to sort a data set and remove duplicate observations at the same time.

The NODUPRECS option in PROC SORT is an easy way to request the removal of the duplicates. However, to guarantee proper removal of all duplicates, you must list every variable in the data set on the BY statement.

The old way: List All Variables in the BY Statement



The original approach is to list all of the variables in the data set in the BY statement, starting with the sort key variables.

```
data SofaData;
  set sashelp.prdsal3;
  where product="SOFA"
    and date between "01mar1998"d and "30apr1998"d
    and state in ("California" "Florida");
  keep State Date Actual;
run;
proc print data=SofaData;
  title "Sofa Data (There is a duplicate for California in Mar98)";
run;
proc sort data=SofaData out=ByStateDate noduprecs;
  by State Date;
run;
proc print data=ByStateDate;
  where State="California";
  title "California remains intact";
run;
  * REMOVE DUPLICATES, SORTING BY ALL VARIABLES (explicit);
proc sort data=SofaData out=ByAllVars noduprecs;
  by State Date Actual;
run;
proc print data=ByAllVars;
  where State="California";
  title "Complete Duplicate is Removed";
run;
```

Please direct all correspondence to:

Warren Repole, 1705 Palm Springs Dr, Vienna VA 22182-2331

sasdinosaur@repole.com

www.repole.com/dinosaur

The new way: Use the `_ALL_` Keyword in the BY Statement (available in SAS Release 82)

An alternate approach is to reference all of the variables in the data set through the `_ALL_` keyword.

```
data SofaData;
  set sashelp.prdsal3;
  where product="SOFA"
    and date between "01mar1998"d and "30apr1998"d
    and state in ("California" "Florida");
  keep State Date Actual;
run;
proc sort data=SofaData out=ByAllVars noduprecs;
  by State Date _ALL_;
run;
proc print data=ByAllVars;
  where State="California";
  title1 "Complete Duplicate is Removed";
run;
```

Advantages of the alternate approach:

- No knowledge of the non-key variable names is required.

Disadvantages of the alternate approach:

- PROC SORT produces an extra note in the SAS log indicating that duplicate variable names in the BY statement were ignored.

Additional documentation for this technique can be found in *Base SAS® 9.2 Procedures Guide*. Cary, NC: SAS Institute Inc.

Visit <http://support.sas.com/documentation/onlinedoc/sas9doc.html> for SAS 9 documentation.

Go to <http://www.repole.com/dinosaur/nodups.html> for the sample code and output for this topic.

These techniques are mentioned in other SAS references and publications:

SAS Usage Note 1566(<http://support.sas.com/kb/1/566.html>)

SUGI Paper 037-30(<http://www2.sas.com/proceedings/sugi30/037-30.pdf>)

Cody's Data Cleaning Techniques Using SAS Software; Ron Cody, 1999 .