

SAS Solutions for Fatal Pitfalls Within the Consumer Expenditure Survey

Jon G. Fields, Ph.D., Northern State University, Aberdeen, South Dakota

ABSTRACT

The data within the Consumer Expenditure Survey (CES), collected by the United States Department of Labor, Bureau of Statistics, and the Census Bureau, provides a plethora of valuable data for researchers. Each year of CES data consists of many individual data sets with differing structures. While thousands of researchers use the CES data for various analyses, many are unaware that the structures within the data sets present serious pitfalls that, if left uncontrolled, produce seriously flawed statistical output that lead to erroneous conclusions and misleading publications.

This paper discusses the problems created by the existence of multiple observations for individual households within some of the CES data sets and illustrates the ramifications of those multiple observations. It also discusses the problems created by the absence of data for individual households in some of the data sets that must be merged with other CES data sets that do contain data for those households. This paper also provides an example of criteria-based selection when deciding which of the multiple observations to include in the actual research project, and it provides SAS examples that sort and select appropriate observations for the actual statistical analysis. And finally, this paper provides a SAS example to correctly merge CES data sets.

INTRODUCTION: THE PITFALLS

A single year of data within the Consumer Expenditure Survey (CES) consists of approximately twenty-five subsets of data. Each subset of data contains information that is logically related, such as family demographics, trips and vacations, appliances, etc. One pitfall within several of the CES data sets is the presence of multiple observations for an individual household within a single data set. For example, if one household reported taking one vacation trip during the survey period, that household would have one observation in the data set. If a second household reported taking two trips during the same period, that household would have two separate observations in the data set. If this condition is not recognized and controlled, the analysis will interpret the data as it were for three different households taking one trip each. See the Figure 1.

<u>Observation</u>	<u>Household #</u>	<u>Expenditure</u>
1	101	\$50
2	103	\$50
3	103	\$50

Figure 1. Dataset illustrating multiple observations.

The data in Figure 1 shows that household 101 reported taking one trip and spending \$50, while household 103 reported taking two trips and spending \$50 on each trip. If the analysis was to determine the mean expenditure for a trip, the correct answer would be \$50 $\{(\$50 + \$50 + \$50) / 3 \text{ trips} = \$50\}$. However, if the analysis was to determine the mean amount that households spend on travel, and if the presence of multiple observations was not controlled, the faulty analysis will read each of the three observations as three separate households and return an incorrect answer of \$50, while the correct answer should be \$75 $\{(\$50 + \$100) / 2 \text{ households} = \$75\}$. The potential for analytical errors increase dramatically as some households may take five, ten, or fifteen qualifying trips during the survey period and spending various amounts of money per trip. The reader should note that, while the above discussion uses trips and expenditures as an example, the pitfall of multiple observations per household occurs in many other subsets of CES data.

A second pitfall within the CES data sets is the total absence of households in some data sets. Data for an individual household may be found in many of these data subsets, but may not be present in other data subsets if the variables in those data sets do not apply to that household. For example, households that have not taken any qualifying trips during the survey period and are not represented in the subset of data related to trips and vacations nor is the fact that they spent \$0 on travel included in the analysis. In Figure 1, while household number 102 is not present in the data set related to trips and vacations since it did not report making a qualified trip during the survey period, it is present in other subsets of the CES data. Depending on the research question being asked, an analysis made without considering these absent households could result in faulty interpretations and publications, and could mislead other researchers attempting to expand on the effort.

Variables of interest for a particular study will likely require accessing variables located in two or more of the CES data sets. Analyses will probably require merging these data sets into a single larger data set. The merge effort introduces more problems if the proper controls are not taken, as the following example illustrates. See Figure 2.

Data Set 1		+	Data Set 2		=	Merged Data Set		
Household	Pet		Household	Pet Food		Household	Pet	Pet Food
101	Cat		101	Cat Food		101	Cat	Cat Food
102	Dog		102	Dog Food		102	Dog	Dog Food
			102	Bird Food		102	Dog	Bird Food

Figure 2. Illustration of incorrect merge results.

Suppose the responses in data set (1) report the kinds of pets that currently live in each household, and that the responses in data set (2) report the kinds of pet foods present in each household. Household 101 reports the presence of a cat (data set 1) and cat food (data set 2). The resulting merged data set for household 101 shows a correct merge outcome. Household 102 reports the presence of a dog (data set 1) and has both dog food and bird food (data set 2) for a bird that is no longer or not yet present in the household.

Notice that household 102 has a single observation in data set (1) and multiple observations in data set (2). When SAS merges data sets (1) and (2) it finds one observation for household 102 in data set (1) and two observations for household 102 in data set (2). SAS duplicates the last observation that it encountered for household 102 within data set (1) and matches that duplicated information with the second observation for household 102 in data set (2). The resulting merged data set contains two complete observations for household 102. While no error messages are produced by SAS, the merged data set incorrectly reports that household 102 has two dogs in the household rather than one, and that one dog is fed dog food and the other is fed bird seeds. Analyses performed on the Merged Data Set would clearly contain critical and fatal errors. The problem with multiple observations for individual households is magnified when several CES data sets are merged and each data set has varying numbers of observations for a given household.

THREE SAS SOLUTIONS FOR MULTIPLE OBSERVATIONS PER HOUSEHOLD

Many analyses using the CES data sets require researchers to merge two or more of the data sets to bring together variables of interest into a single analysis data set. Before any merge process takes place, those households having multiple observations must be manipulated to eliminate some of the multiple observations, or to collapse the multiple observations into a single observation.

The first step is determine which of the data sets of interest actually have multiple observations per household (households are identified by the variable name 'NEWID'). This can be easily done by performing a visual review of the data in each of the data sets with the following commands.

```
DATA DataSet1;
PROC SORT; BY NewID
PROC PRINT DATA= DataSet1 (obs=40);
```

These commands sort the households by their ID numbers and print the first forty observations of the data set. The number forty is arbitrary, but should be sufficient to determine whether or not multiple observations per household are present.

SOLUTION 1:

The elimination of multiple observations for a given household must be performed in each data set of interest before the data sets are merged. One solution is to simply accept the first observation that SAS encounters and to drop or delete any subsequent observations that it encounters for that same household. This can be done with a SORT procedure.

```
PROC SORT NoDupKey; BY NEWID;
```

The above SORT procedure sorts the data set, retains the first occurrence of NEWID that it encounters, and drops all subsequent observations having the same NEWID. (PROC SORT NODUPKEY looks only for duplicates of the variable specified in the BY statement, while PROC SORT NODUP and PROC SORT NODUPPLICATES look at *all* the values in *all* the variables of an observation to look for exact duplicates of an *entire* observation).

SOLUTION 2:

While Solution 1 is simple to do, the first observation per household that SAS encounters and retains using this NODUPKEY procedure may not be the 'best' observation to retain for the particular study at hand. Researchers should establish criteria to identify the most desirable observations to retain. For example, some households may travel several times during a particular year. One or more of these trips may be taken on a 'package' deal and other trips may be taken on a 'non-package' basis. The data set containing travel data includes both 'package' trips and 'non-package' trips. Let's assume that an analysis of 'package' versus 'non-package' trips is of interest. Since traveling on a 'package' deal occurs much less frequently than traveling on a 'non-package' deal, retaining the observations that include 'package' trips should be given priority. The criteria a researcher may use in this situation might be to give 'package' trips priority for retention, retain the 'package' trip that occurred most recently, and retain the trips that were the longest in duration.

```

PROC SORT;
  BY NEWID                      /* Group household observations together */
  PKGTRIP                      /* Package trips (1=Yes) get 1st priority */
  descending INTV_NUM          /* Most recent interview gets 2nd priority */
  descending TRIPNTSQ;         /* Trips with most nights get 3rd priority */

PROC SORT NODUPKEY; BY NEWID; /* Retains only the first observation
                               per household encountered */

```

In the above example, the first SORT procedure sorts and arranges the data based on the priority criteria established by researcher. The second SORT procedure retains the first trip it encounters (with the highest priority as defined by the criteria) and deletes any other trip observations (multiple observations) taken by the same households and having lower priority. The resulting data set contains only one trip observation per household. Figure 3 helps to illustrate what has happened. The observations in *italics* help to identify the desired observations to be retained based on the established criteria.

<u>Before any sort</u>			
Household (NewID)	(TripType)	Interview # (Intv Num)	# Nights (TripNtsQ)
101	non-pkg	3	4
102	non-pkg	3	4
<i>101</i>	<i>non-pkg</i>	4	3
<i>102</i>	<i>package</i>	4	4
102	non-pkg	5	3
<u>After first sort procedure</u>			
<i>101</i>	<i>non-pkg</i>	4	3
101	non-pkg	3	4
<i>102</i>	<i>package</i>	4	4
102	non-pkg	5	3
102	non-pkg	3	4
<u>After second sort procedure</u>			
<i>101</i>	<i>non-pkg</i>	4	3
<i>102</i>	<i>package</i>	4	4

Figure 3. Results of Solution 2 SORT procedures.

INTERVENING SOLUTION 3:

In some cases, direct use of Solutions 1 or 2 may be sufficient for a particular study. A drawback of going directly to one of these solutions, however, is that some potentially useful data is lost when observations are dropped during the DoDupKey step. For example, if the researcher would like to determine the total amount that a household spent on travel over the course of several trips (over several observations), this information cannot be determined from the subsets of data that are produced by applying Solutions 1 or 2 too soon. An intervening procedure that captures information from multiple observations per household is needed.

The CES data records the cost of lodging during a trip with the variable name LDGCOSTX. Let's suppose that household 102 took three trips and paid differing amounts for lodging for each trip. The data set would contain three separate observations (multiple observations) for this household with each observation providing important trip and cost data. See Figure 4.

<u>DataSet1</u>		
Household (NewID)	Lodging Cost (LdgCostX)	(Other variables...→)
102	50	
102	100	
102	150	

Figure 4. Original trip data for household 102

The objectives here are to sum the total lodging expenditures for household 102 and to retain that sum in the single observation that will be retained and used in the final analysis. Special SUM and MERGE procedures accomplish the first of these objectives.

```
PROC MEANS SUM NOPRINT;
  BY NEWID;                      /* Sum by common household number NEWID */
  VAR LdgCostX;                  /* Sums the LdgCostX variable */
  OUTPUT OUT= DataSet2          /* Creates an output data set called DATASET2 */
  SUM= Cost;                     /* The new output variable in DATASET2
                                containing the sum is called COST. */
```

This PROC MEANS SUM procedure produces a single observation for household 102 in DataSet2 with a COST value of \$300. See Figure 5

<u>DataSet2</u>	
Household	
(NewID)	(Cost)
102	300

Figure 5. Result of PROC MEANS SUM procedure.

A MERGE step to combine DataSet1 and DataSet2 might be written as follows, and would produce the results in Figure 6. Notice that the MERGE step in SAS has duplicated the COST value in DataSet2 and has matched that duplicated value to the second and third observations in DataSet1.

```
DATA DataSet3;
  MERGE DataSet1 DataSet2;
  BY NEWID;
```

<u>DataSet3</u>			
Household	Lodging Cost		
(NewID)	(LdgCostX)	(Cost)	(Other variables...→)
102	50	300	
102	100	300	
102	150	300	

Figure 6. Results from merging DataSet1 and DataSet2 into the new DataSet3.

Applying the SORT procedures from either Solution 1 or Solution 2 described above to DataSet3 will result in a new data set that contains the composite information from across multiple observations for each household, in this case, a COST value of \$300. Now, no matter which criteria-based observation is selected and retained in the SORT procedures, each household's composite information will be included in the final analysis data set. IF/THEN statements and other commands can be used to count the number of observations per household or to capture other composite information.

SAS SOLUTION FOR MERGING DATA SETS WITH MISSING OBSERVATIONS

SAS provides several methods to combine two or more data sets into a single larger data set including match-merging, one-to-one merging, one-to-one reading, concatenating, interleaving, and updating. (See *SAS Language Statements and Rules of the SAS Language* for more information.) This author recommends that the match-merging method be used when combining CES data sets.

Each year of CES data contains approximately twenty-five data sets with each containing data on thousands of households. Some of the files are for specific calendar quarters of the year, or cover specific topics like trips/vacations, vehicles, and appliances. For most research efforts, it is necessary to merge several CES data sets into a single analysis data set.

Households that do not have data relevant to the special topic data sets are not represented (do not have observations) in those files. That means a particular household may have data in some of the files and be 'missing' in other data files. In an effort to keep the merged analysis data set to a more manageable size, it is recommended that specific variables be retained with appropriate KEEP statements and to eliminate household that have 'missing' data by using appropriate MERGE statements. The following discussion focuses on the MERGE procedures.

Each data set that is to be merged must first be reduced to a single observation per household (NEWID) as described earlier, sorted, and then merged. The following example illustrates the sort and merge steps.

```

DATA CESData1;
  PROC SORT; BY NEWID;

DATA CESData2;
  PROC SORT; BY NEWID;

DATA MergedData;
  MERGE CESData1 (IN= inSet1)
        CESData2 (IN= inSet2)
  BY NEWID;
  IF inSet1=1 and inSet2=1;

```

The IN= option creates the new variables inSet1 and inSet2. If a household has a NEWID in both the CESData1 and in CESData2 data sets, then the values for the inSet1 and inSet2 variables are set equal to one (1). If an observation for a household is not found in both the Set1 and Set2 data sets, then the values for the inSet1 and inSet2 variables are set equal to zero (0). The IF statement selects (keeps) only those household (NEWID) observations in which the inSet1 and inSet2 variables are both equal to '1'. With this process, only those observations with data in both data sets are retained in the analysis data set.

CONCLUSION

The collection of data sets that make up each year's Consumer Expenditure Survey (CES) offers researchers many opportunities to perform large-scale and detailed analyses on a multitude of research topics. Using the CES properly requires researchers to become intimately familiar with the structure and contents of the data sets of interest. Failure to do so, and not to recognize and control for potential pitfalls will undoubtedly result in fatal errors in the research design, the statistical output that is generated, and in any conclusions drawn from those flawed efforts.

While this paper concentrated on the pitfalls of multiple observations and merging often encountered in the CES data, other data selection pitfalls also exist. These include the duplication of some data in consecutive years of CES data sets, households that skip one or more of the quarterly interviews, households that have some of their quarterly data contained in one year of CES data and some of their quarterly data contained in another year of CES data, all of which leave gaps in the data and could easily lead to inaccurate statistical output. This author recommends that researchers spend large amounts of time becoming familiar with the CES data, consider the analytical ramifications of what they learn, and to take appropriate steps in order to produce quality output.

CONTACT INFORMATION

Jon G. Fields, Ph.D.
 Northern State University
 1200 South Jay Street
 Campus Box 689
 Aberdeen, SD 57401
 Phone: (605) 626-3003
 Fax: (605) 626-2572
 e-Mail: fieldsj@northern.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.