THE POWER TO KNOW.

SAS/STAT® 12.3 User's Guide

Special SAS Data Sets (Chapter)

This document is an individual chapter from *SAS/STAT® 12.3 User's Guide*.

The correct bibliographic citation for the complete manual is as follows: SAS Institute Inc. 2013. *SAS/STAT® 12.3 User's Guide*. Cary, NC: SAS Institute Inc.

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

July 2013

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at **support.sas.com/bookstore** or call 1-800-727-3228.

Appendix A
Special SAS Data Sets

Contents

Introduction to Special SAS Data Sets

All SAS/STAT procedures create SAS data sets. Any table generated by a procedure can be saved to a data set by using the Output Delivery System (ODS), and many procedures also have syntax that enables you to save other statistics to data sets. Some of these data sets are organized according to certain conventions so that they can be read by a SAS/STAT procedure for further analysis. Such specially organized data sets are recognized by the TYPE= data set attribute.

The CORR procedure (see the *Base SAS Procedures Guide: Statistical Procedures*), for example, can create a data set with the attribute TYPE=CORR containing a correlation matrix. This TYPE=CORR data set can

be read by the REG or FACTOR procedure, among others. If the original data set is large, using a special SAS data set in this way can save computer time by avoiding the recomputation of the correlation matrix in subsequent analyses.

PROC REG, for example, can create a TYPE=EST data set containing estimated regression coefficients. If you need to make predictions for new observations, you can use the SCORE procedure to read both the TYPE=EST data set and a data set containing the new observations. PROC SCORE can then compute predicted values or residuals without repeating the entire regression analysis. See Chapter 82, "The SCORE Procedure," for an example.

A special SAS data set might contain different kinds of statistics. A special variable called _TYPE_ is used to distinguish the various statistics. For example, in a TYPE=CORR data set, an observation in which _TYPE_='MEAN' contains the means of the variables in the analysis, and an observation in which _TYPE_='STD' contains the standard deviations. Correlations appear in observations with _TYPE_='CORR'. Another special variable, _NAME_, is needed to identify the row of the correlation matrix. Thus, the correlation between variables X and Y is given by the value of the variable X in the observation for which _TYPE_='CORR' and _NAME_='Y', or by the value of the variable Y in the observation for which _TYPE_='CORR' and _NAME_='X'.

The special data sets created by SAS/STAT procedures can generally be used directly by other procedures without modification. However, if you create an output data set with PROC CORR and use the NOCORR option to omit the correlation matrix from the OUT= data set, you need to set the TYPE= option either in parentheses following the OUT= data set name in the PROC CORR statement or in parentheses following the DATA= option in any other procedure that recognizes the special TYPE= attribute. In either case, the TYPE= option should be set to COV, CSSCP, or SSCP according to what type of matrix is stored in the data set and what data set types are accepted as input by the other procedures you plan to use. If you do not follow these steps and you use the TYPE=CORR data set with no correlation matrix as input to another procedure, the procedure might issue an error message indicating that the correlation matrix is missing from the data set.

You can create special SAS data sets directly in a DATA step by specifying the TYPE= option in parentheses after the data set name in the DATA statement. See "Example A.2: Creating a TYPE=CORR Data Set in a DATA Step" on page 8696 for an example. If you use a DATA step with a SET statement to modify a special SAS data set, you must specify the TYPE= option in the DATA statement. The TYPE= attribute of the data set in the SET statement is *not* automatically copied to the data set being created. You can determine the TYPE= attribute of a data set by using the CONTENTS procedure (see "Example A.1: A TYPE=CORR Data Set Produced by PROC CORR" on page 8695 and the *Base SAS Procedures Guide* for details).

Table A.1 summarizes the TYPE= data sets that can be used as input to SAS/STAT procedures. Table A.2 summarizes the TYPE= data sets that are created by SAS/STAT procedures and the statements each procedure uses to create its special output data sets. Most procedures accept ordinary SAS data sets and create ordinary output SAS data sets with no TYPE= specification in addition to the special data sets shown in the tables. When you specify a data set with a type that the procedure does not recognize, the procedure prints an error message and stops executing.

Table A.1 SAS/STAT Procedures That Accept Special Input Data Sets Types

Procedure	Special TYPE= Data Sets Accepted
ACECLUS	ACE, CORR, COV, SSCP, UCORR, UCOV
BOXPLOT	BOXPLOT, CHARTSUM
CALIS	CALISMDL, CORR, COV, FACTOR, SSCP, UCORR, UCOV, WEIGHT
CANDISC	CORR, COV, SSCP, CSSCP
CATMOD	EST
CLUSTER	DISTANCE
DISCRIM	CORR, COV, SSCP, CSSCP, LINEAR, QUAD, MIXED
FACTOR	ACE, CORR, COV, FACTOR, SSCP, UCORR, UCOV
LIFEREG	EST
LOGISTIC	EST LOGISMOD
MI	EST, COV, CORR
MIANALYZE	EST, COV, CORR
MODECLUS	DISTANCE
PHREG	EST
PRINCOMP	ACE, CORR, COV, EST, FACTOR, SSCP, UCORR, UCOV
PROBIT	EST
QUANTREG	EST
REG	CORR, COV, SSCP, UCORR, UCOV
ROBUSTREG	EST
SCORE	SCORE= data set can be of any type
SIMNORM	CORR, COV
SURVEYLOGISTIC	EST
STEPDISC	CORR, COV, SSCP, CSSCP
TREE	TREE
VARCLUS	CORR, COV, FACTOR, SSCP, UCORR, UCOV

Table A.2 SAS/STAT Procedures That Create Special Output Data Set Types

Procedure	TYPE=	Statement and Option Required
ACECLUS	ACE	PROC ACECLUS OUTSTAT=
BOXPLOT	BOXPLOT	PLOT / OUTBOX=
	CHARTSUM	PLOT / OUTHISTORY=
CALIS	CALISFIT	PROC CALIS OUTFIT=
	CALISMDL	PROC CALIS OUTMODEL=
	CORR	PROC CALIS CORR OUTSTAT=
	COV	PROC CALIS OUTSTAT=
	EST	PROC CALIS OUTEST=
	WEIGHT	PROC CALIS OUTWGT=
CANCORR	CORR	PROC CANCORR OUTSTAT=
	UCORR	PROC CANCORR NOINT OUTSTAT=
CANDISC	CORR	PROC CANDISC OUTSTAT=
CATMOD	EST	RESPONSE / OUTEST=

Table A.2 *continued*

Procedure	TYPE=	Statement and Option Required
CLUSTER	TREE	PROC CLUSTER OUTTREE=
DISCRIM	LINEAR	PROC DISCRIM POOL=YES OUTSTAT=
	QUAD	PROC DISCRIM POOL=NO OUTSTAT=
	MIXED	PROC DISCRIM POOL=TEST OUTSTAT=
	CORR	PROC DISCRIM METHOD=NPAR OUTSTAT=
DISTANCE	DISTANCE	PROC DISTANCE METHOD=*distance-method* OUT=
	SIMILAR	PROC DISTANCE METHOD=*similarity-method* OUT=
FACTOR	FACTOR	PROC FACTOR OUTSTAT=
LIFEREG	EST	PROC LIFEREG OUTEST=
LOGISTIC	EST	PROC LOGISTIC OUTEST=
	LOGISMOD	PROC LOGISTIC OUTMODEL=
MI	COV	EM OUTEM=
	COV	EM OUTITER=
	COV	MCMC OUTITER=
	EST	MCMC OUTEST=
NLIN	EST	PROC NLIN OUTEST=
ORTHOREG	EST	PROC ORTHOREG OUTEST=
PHREG	EST	PROC PHREG OUTEST=
PRINCOMP	CORR	PROC PRINCOMP OUTSTAT=
	COV	PROC PRINCOMP COV OUTSTAT=
	UCORR	PROC PRINCOMP NOINT OUTSTAT=
	UCOV	PROC PRINCOMP NOINT COV OUTSTAT=
PROBIT	EST	PROC PROBIT OUTEST=
QUANTREG	EST	PROC QUANTREG OUTEST=
REG	EST	PROC REG OUTEST=
	SSCP	PROC REG OUTSSCP=
ROBUSTREG	EST	PROC ROBUSTREG OUTEST=
VARCLUS	CORR	PROC VARCLUS OUTSTAT=
	UCORR	PROC VARCLUS NOINT OUTSTAT=
	TREE	PROC VARCLUS OUTTREE=

Special SAS Data Sets

TYPE=ACE Data Sets

A TYPE=ACE data set is created by the ACECLUS procedure, and it contains the approximate within-cluster covariance estimate, as well as eigenvalues and eigenvectors from a canonical analysis, among other statistics. It can be used as input to the ACECLUS procedure to initialize another execution of PROC ACECLUS. It can also be used to compute canonical variable scores with PROC SCORE and as input to PROC FACTOR, specifying METHOD=SCORE, to rotate the canonical variables. See Chapter 23, "The ACECLUS Procedure," for details.

TYPE=BOXPLOT Data Sets

A TYPE=BOXPLOT data set is created by and used by the BOXPLOT procedure. The data set contains the group summary statistics and outlier values required for constructing a schematic box plot. Each observation in a TYPE=BOXPLOT data set records the value of a single feature of one group's box-and-whiskers plot, such as its mean. Consequently, a TYPE=BOXPLOT data set contains multiple observations per group. These must appear consecutively in the data set. The _TYPE_ variable identifies the feature whose value is recorded in a given observation. _TYPE_ values of 'N', 'MIN', 'Q1', 'MEDIAN', 'MEAN', 'Q3', and 'MAX 'are required for each group. See Chapter 26, "The BOXPLOT Procedure," for details.

TYPE=CALISFIT Data Sets

PROC CALIS creates a TYPE=CALISFIT data set. This data set contains the names of the model fit indices and their values. A TYPE=CALISFIT data set is intended to save all the fit index values for future use, especially when the customized fit summary table shows only a small number of fit indices. See Chapter 27, "The CALIS Procedure," for details.

TYPE=CALISMDL Data Sets

PROC CALIS creates and accepts as input a TYPE=CALISMDL data set. This data set contains the model specification and the computed parameter estimates. A TYPE=CALISMDL data set is intended to be reused as an input data set to specify good initial values in subsequent analyses by PROC CALIS. See Chapter 27, "The CALIS Procedure," for details.

TYPE=CHARTSUM Data Sets

A TYPE=CHARTSUM data set is created by and used by the BOXPLOT procedure. The data set contains group summary statistics associated with box-and-whiskers plots. See Chapter 26, "The BOXPLOT Procedure," for details.

TYPE=CORR Data Sets

A TYPE=CORR data set usually contains a correlation matrix and possibly other statistics including means, standard deviations, and the number of observations in the original SAS data set from which the correlation matrix was computed. Using PROC CORR with an output data set option (OUTP=, OUTS=, OUTK=, OUTH=, or OUT=) produces a TYPE=CORR data set. (For a complete description of the CORR procedure, see the *Base SAS Procedures Guide: Statistical Procedures*.) The CALIS, CANCORR, CANDISC, DISCRIM, PRINCOMP, and VARCLUS procedures can also create a TYPE=CORR data set with additional statistics (the CORR option is needed in PROC CALIS). A TYPE=CORR data set containing a correlation matrix can be used as input for the ACECLUS, CALIS, CANCORR, CANDISC, DISCRIM, FACTOR, PRINCOMP, REG, SCORE, STEPDISC, and VARCLUS procedures. The variables in a TYPE=CORR data set are as follows:

- the BY variable or variables, if a BY statement is used with the procedure
- _TYPE_, a character variable of length eight with values identifying the type of statistic in each observation, such as 'MEAN', 'STD', 'N', and 'CORR'
- _NAME_, a character variable with values identifying the variable with which a given row of the correlation matrix is associated
- other variables that were analyzed by the CORR procedure or other procedures

The usual values of the _TYPE_ variable are as follows:

TYPE	Contents
MEAN	mean of each variable analyzed
STD	standard deviation of each variable
N	number of observations used in the analysis. PROC CORR records the number of nonmissing values for each variable unless the NOMISS option is used. If the NOMISS option is specified, or if the CALIS, CANCORR, CANDISC, PRINCOMP, or VARCLUS procedure is used to create the data set, observations with one or more missing values are omitted from the analysis, so this value is the same for each variable and provides the number of observations with no missing values. If a FREQ statement is used with the procedure that creates the data set, the number of observations is the sum of the relevant values of the variable in the FREQ statement. Procedures that read a TYPE=CORR data set use the smallest value in the observation with _TYPE_='N' as the number of observations in the analysis.
SUMWGT	sum of the observation weights if a WEIGHT statement is used with the procedure that creates the data set. The values are determined analogously to those of the _TYPE_='N' observation.
CORR	correlations with the variable named by the _NAME_ variable

There might be additional observations in a TYPE=CORR data set depending on the particular procedure and options used.

If you create a TYPE=CORR data set yourself, the data set need not contain the observations with _TYPE_='MEAN', 'STD', 'N', or 'SUMWGT', unless you intend to use one of the discriminant procedures. Procedures assume that all of the means are 0.0 and that the standard deviations are 1.0 if this information is not in the TYPE=CORR data set. If _TYPE_='N' does not appear, most procedures assume that the number of observations is 10,000; significance tests and other statistics that depend on the number

of observations are, of course, meaningless. In the CALIS and CANCORR procedures, you can use the EDF= option instead of including a _TYPE_='N' observation.

A correlation matrix is symmetric; that is, the correlation between X and Y is the same as the correlation between Y and X. The CALIS, CANCORR, CANDISC, CORR, DISCRIM, PRINCOMP, and VARCLUS procedures output the entire correlation matrix. If you create the data set yourself, you need to include only one of the two occurrences of the correlation between two variables; the other can be given a missing value.

If you create a TYPE=CORR data set yourself, the _TYPE_ and _NAME_ variables are not necessary except for use with the discriminant procedures and PROC SCORE. If there is no _TYPE_ variable, then all observations are assumed to contain correlations. If there is no _NAME_ variable, the first observation is assumed to correspond to the first variable in the analysis, the second observation to the second variable, and so on. However, if you omit the _NAME_ variable, you will not be able to analyze arbitrary subsets of the variables or list the variables in a VAR or MODEL statement in a different order.

Example A.1: A TYPE=CORR Data Set Produced by PROC CORR

See Figure A.1 for an example of a TYPE=CORR data set produced by the following SAS statements. Figure A.2 displays partial output from PROC CONTENTS, which indicates that the "Data Set Type" is 'CORR'.

```
title 'Five Socioeconomic Variables';
title2 'Harman (1976), Modern Factor Analysis, Third Edition';

data SocEcon;
   input Pop School Employ Services House;
   datalines;
5700     12.8      2500      270       25000
1000     10.9      600       10        10000
3400     8.8       1000      10        9000
3800     13.6      1700      140       25000
4000     12.8      1600      140       25000
8200     8.3       2600      60        12000
1200     11.4      400       10        16000
9100     11.5      3300      60        14000
9900     12.5      3400      180       18000
9600     13.7      3600      390       25000
9600     9.6       3300      80        12000
9400     11.4      4000      100       13000
;

proc corr noprint out=corrcorr;
run;

proc print data=corrcorr;
run;

proc contents data=corrcorr;
run;
```

Figure A.1 A TYPE=CORR Data Set Produced by PROC CORR

Five Socioeconomic Variables
Harman (1976), Modern Factor Analysis, Third Edition

Obs	_TYPE_	_NAME_	Pop	School	Employ	Services	House
1	MEAN		6241.67	11.4417	2333.33	120.833	17000.00
2	STD		3439.99	1.7865	1241.21	114.928	6367.53
3	N		12.00	12.0000	12.00	12.000	12.00
4	CORR	Pop	1.00	0.0098	0.97	0.439	0.02
5	CORR	School	0.01	1.0000	0.15	0.691	0.86
6	CORR	Employ	0.97	0.1543	1.00	0.515	0.12
7	CORR	Services	0.44	0.6914	0.51	1.000	0.78
8	CORR	House	0.02	0.8631	0.12	0.778	1.00

Figure A.2 Contents of a TYPE=CORR Data Set

Five Socioeconomic Variables
Harman (1976), Modern Factor Analysis, Third Edition

The CONTENTS Procedure

Data Set Name	WORK.CORRCORR	Observations	8
Member Type	DATA	Variables	7
Engine	SASE7	Indexes	0
Created	DDMMMYY:00:00:00	Observation Length	56
Last Modified	DDMMMYY:00:00:00	Deleted Observations	0
Protection		Compressed	NO
Data Set Type	CORR	Sorted	NO
Label	Pearson Correlation Matrix		
Data Representation	Native		
Encoding	Session		

Example A.2: Creating a TYPE=CORR Data Set in a DATA Step

This example creates a TYPE=CORR data set by reading a correlation matrix in a DATA step. Figure A.3 shows the resulting data set.

```
title 'Five Socioeconomic Variables';

data datacorr(type=corr);
   infile cards missover;
   _type_='corr';
   input _Name_ $ Pop School Employ Services House;
   datalines;
Pop         1.00000
School      0.00975   1.00000
Employ      0.97245   0.15428   1.00000
Services    0.43887   0.69141   0.51472   1.00000
House       0.02241   0.86307   0.12193   0.77765   1.00000
;
```

```
proc print data=datacorr;
run;
```

Figure A.3 A TYPE=CORR Data Set Created by a DATA Step

Five Socioeconomic Variables

OBS	_type_	_Name_	Pop	School	Employ	Services	House
1	corr	Pop	1.00000	.	.	.	.
2	corr	School	0.00975	1.00000	.	.	.
3	corr	Employ	0.97245	0.15428	1.00000	.	.
4	corr	Services	0.43887	0.69141	0.51472	1.00000	.
5	corr	House	0.02241	0.86307	0.12193	0.77765	1

TYPE=COV Data Sets

A TYPE=COV data set is similar to a TYPE=CORR data set except that it has _TYPE_='COV' observations containing covariances instead of or in addition to _TYPE_='CORR' observations containing correlations. The CALIS and PRINCOMP procedures create a TYPE=COV data set (the COV option is needed in PROC PRINCOMP). You can also create a TYPE=COV data set by using PROC CORR with the COV and NOCORR options and specifying the data set option TYPE=COV in parentheses following the name of the output data set. You can use only the OUTP= or OUT= option to create a TYPE=COV data set with PROC CORR. Another way to create a TYPE=COV data set is to read a covariance matrix in a data set, in the same manner as shown in "Example A.2: Creating a TYPE=CORR Data Set in a DATA Step" on page 8696 for a TYPE=CORR data set. TYPE=COV data sets are used by the same procedures that use TYPE=CORR data sets.

TYPE=CSSCP Data Sets

A TYPE=CSSCP data set contains a corrected sum of squares and crossproducts (CSSCP) matrix. TYPE=CSSCP data sets are created by using the CORR procedure with the CSSCP option and specifying the data set option TYPE=CSSCP in parentheses following the name of the OUTP= or OUT= data set. You can also create TYPE=CSSCP data sets in a DATA step; in this case, TYPE=CSSCP must be specified as a data set option. The variables in a TYPE=CSSCP data set are the same as those found in a TYPE=SSCP data set, except that there is not a variable called Intercept or a row with _NAME_='Intercept'. TYPE=CSSCP data sets are read by only the CANDISC, DISCRIM, and STEPDISC procedures. Formulas useful for illustrating differences between corrected and uncorrected matrices in some special SAS data sets are shown in the section "Definitional Formulas" on page 8702.

TYPE=DISTANCE Data Sets

PROC DISTANCE creates a TYPE=DISTANCE or TYPE=SIMILAR data set, depending on the METHOD= option. TYPE=DISTANCE can be used as an input data set to PROC MODECLUS or

PROC CLUSTER, but TYPE=SIMILAR cannot be used as an input to any procedures. The proximity measures are stored as a lower triangular matrix or a square matrix in the OUT= data set (depending on the SHAPE= option). See Chapter 34, "The DISTANCE Procedure," for details. You can also create a TYPE=DISTANCE data set in a DATA step by reading or computing a lower triangular or symmetric matrix of dissimilarity values, such as a chart of mileage between cities. The number of observations must be equal to the number of variables used in the analysis. This type of data set is used as input by the CLUSTER and MODECLUS procedures. PROC CLUSTER ignores the upper triangular portion of a TYPE=DISTANCE data set and assumes that all main diagonal values are zero, even if they are missing. PROC MODECLUS uses the entire distance matrix and does not require the matrix to be symmetric. See Chapter 31, "The CLUSTER Procedure," and Chapter 60, "The MODECLUS Procedure," for examples and details.

TYPE=EST Data Sets

A TYPE=EST data set contains parameter estimates. The CALIS, CATMOD, LIFEREG, LOGISTIC, NLIN, ORTHOREG, PHREG, PROBIT, and REG procedures create TYPE=EST data sets when the OUTEST= option is specified. A TYPE=EST data set produced by PROC LIFEREG, PROC ORTHOREG, or PROC REG can be used with PROC SCORE to compute residuals or predicted values. The variables in a TYPE=EST data set include the following:

- the BY variables, if a BY statement is used
- _TYPE_, a character variable of length eight, that indicates the type of estimate. The values depend on which procedure created the data set. Usually a value of 'PARM' or 'PARMS' indicates estimated regression coefficients, and a value of 'COV' or 'COVB' indicates estimated covariances of the parameter estimates. Some procedures, such as PROC NLIN, have other values of _TYPE_ for special purposes.
- _NAME_, a character variable that contains the values of the names of the rows of the covariance matrix when the procedure outputs the covariance matrix of the parameter estimates
- variables that contain the parameter estimates, usually the same variables that appear in the VAR statement or in any MODEL statement. See Chapter 27, "The CALIS Procedure," Chapter 30, "The CATMOD Procedure," and Chapter 63, "The NLIN Procedure," for details on the variable names used in output data sets created by those procedures.

Other variables can be included depending on the particular procedure and options used.

Example A.3: A TYPE=EST Data Set Produced by PROC REG

Figure A.4 shows the TYPE=EST data set produced by the following statements:

```
proc reg data=SocEcon outest=regest covout;
   full:   model house=pop school employ services / noprint;
   empser: model house=employ services / noprint;
run; quit;
```

```
proc print data=regest;
run;
```

Figure A.4 A TYPE=EST Data Set Produced by PROC REG

Five Socioeconomic Variables

OBS	_MODEL_	_TYPE_	_NAME_	_DEPVAR_	_RMSE_	Intercept
1	full	PARMS		House	3122.03	-8074.21
2	full	COV	Intercept	House	3122.03	109408014.44
3	full	COV	Pop	House	3122.03	-9157.04
4	full	COV	School	House	3122.03	-9784744.54
5	full	COV	Employ	House	3122.03	20612.49
6	full	COV	Services	House	3122.03	102764.89
7	empser	PARMS		House	3789.96	15021.71
8	empser	COV	Intercept	House	3789.96	5824096.19
9	empser	COV	Employ	House	3789.96	-1915.99
10	empser	COV	Services	House	3789.96	-1294.94

OBS	Pop	School	Employ	Services	House
1	0.65	2140.10	-2.92	27.81	-1
2	-9157.04	-9784744.54	20612.49	102764.89	.
3	2.32	852.86	-6.20	-5.20	.
4	852.86	907886.36	-2042.24	-9608.59	.
5	-6.20	-2042.24	17.44	6.50	.
6	-5.20	-9608.59	6.50	202.56	.
7	.	.	-1.94	53.88	-1
8	.	.	-1915.99	-1294.94	.
9	.	.	1.15	-6.41	.
10	.	.	-6.41	134.49	.

TYPE=FACTOR Data Sets

A TYPE=FACTOR data set is created by PROC FACTOR when the OUTSTAT= option is specified. The CALIS, CANCORR, FACTOR, PRINCOMP, SCORE, and VARCLUS procedures can use TYPE=FACTOR data sets as input. The variables are the same as in a TYPE=CORR data set. The statistics include means, standard deviations, sample size, correlations, eigenvalues, eigenvectors, factor patterns, residual correlations, scoring coefficients, and others depending on the options specified. See Chapter 35, "The FACTOR Procedure," for details. When the NOINT option is used with the OUTSTAT= option in PROC FACTOR, the value of the _TYPE_ variable is set to 'USCORE' instead of 'SCORE' to indicate that the scoring coefficients have not been corrected for the mean. If this data set is used with PROC SCORE, the value of the _TYPE_ variable tells PROC SCORE whether or not to subtract the mean from the scoring coefficients.

TYPE=LINEAR Data Sets

A TYPE=LINEAR data set contains the coefficients of a linear function of the variables in observations with _TYPE_='LINEAR'. PROC DISCRIM stores linear discriminant function coefficients in a TYPE=LINEAR data set when you specify METHOD=NORMAL (the default method), POOL=YES, and an OUTSTAT= data set; the data set can be used in a subsequent invocation of PROC DISCRIM to classify additional observations. Many other statistics can be included depending on the options used. See Chapter 33, "The DISCRIM Procedure," for details.

TYPE=LOGISMOD Data Sets

A TYPE=LOGISMOD data set contains information about a logistic regression model fit by PROC LOGISTIC. PROC LOGISTIC both creates and reads TYPE=LOGISMOD data sets. See Chapter 54, "The LOGISTIC Procedure," for details.

TYPE=MIXED Data Sets

A TYPE=MIXED data set contains coefficients of either a linear or a quadratic function, or both if there are BY groups. PROC DISCRIM produces a TYPE=MIXED data set when you specify METHOD=NORMAL (the default method), POOL=TEST, and an OUTSTAT= data set. See Chapter 33, "The DISCRIM Procedure," for details.

TYPE=QUAD Data Sets

A TYPE=QUAD data set contains the coefficients of a quadratic function of the variables in observations with _TYPE_='QUAD'. PROC DISCRIM stores quadratic discriminant function coefficients in a TYPE=QUAD data set when you specify METHOD=NORMAL (the default method), POOL=NO, and an OUTSTAT= data set; the data set can be used in a subsequent invocation of PROC DISCRIM to classify additional observations. Many other statistics can be included depending on the options used. See Chapter 33, "The DISCRIM Procedure," for details.

TYPE=SSCP Data Sets

A TYPE=SSCP data set contains an uncorrected sum of squares and crossproducts (SSCP) matrix. TYPE=SSCP data sets are produced by PROC REG when the OUTSSCP= option is specified in the PROC REG statement. You can also create a TYPE=SSCP data set by using PROC CORR with the SSCP option and specifying the data set option TYPE=SSCP in parentheses following the name of the OUTP= or OUT= data set. You can also create TYPE=SSCP data sets in a DATA step; in this case, TYPE=SSCP must be specified as a data set option.

The variables in a TYPE=SSCP data set include those found in a TYPE=CORR data set. In addition, there is a variable called Intercept that contains crossproducts for the intercept (sums of the variables). The

SSCP matrix is stored in observations with _TYPE_='SSCP', including a row with _NAME_='Intercept'. PROC REG also outputs an observation with _TYPE_='N'. PROC CORR includes observations with _TYPE_='MEAN' and _TYPE_='STD' as well. TYPE=SSCP data sets are used by the same procedures that use TYPE=CORR data sets.

Example A.4: A TYPE=SSCP Data Set Produced by PROC REG

The following statements create a TYPE=SSCP data set from the SocEcon input data set created in "Example A.1: A TYPE=CORR Data Set Produced by PROC CORR" on page 8695:

```
proc reg data=SocEcon outsscp=regsscp;
   model house=pop school employ services / noprint;
run; quit;

proc print data=regsscp;
run;
```

The data set is created by PROC REG and is displayed in Figure A.5.

Figure A.5 A TYPE=SSCP Data Set Produced by PROC REG

Five Socioeconomic Variables

OBS	_TYPE_	_NAME_	Intercept	Pop	School	Employ	Services	House
1	SSCP	Intercept	12.0	74900	137.30	28000	1450	204000
2	SSCP	Pop	74900.0	597670000	857640.00	220440000	10959000	1278700000
3	SSCP	School	137.3	857640	1606.05	324130	18152	2442100
4	SSCP	Employ	28000.0	220440000	324130.00	82280000	4191000	486600000
5	SSCP	Services	1450.0	10959000	18152.00	4191000	320500	30910000
6	SSCP	House	204000.0	1278700000	2442100.00	486600000	30910000	3914000000
7	N		12.0	12	12.00	12	12	12

TYPE=TREE Data Sets

Some clustering procedures produce TYPE=TREE data sets. For example, in PROC CLUSTER, a TYPE=TREE data set contains one observation for each observation in the input data set, plus one observation for each cluster of two or more observations (that is, one observation for each node of the cluster tree). The total number of output observations is usually $2n - 1$, where n is the number of input observations. The density methods might produce fewer output observations when the number of clusters cannot be reduced to one.

In PROC VARCLUS, the OUTTREE= data set contains one observation for each variable clustered plus one observation for each cluster of two or more variables—that is, one observation for each node of the cluster tree. The total number of output observations is between n and $2n - 1$, where n is the number of variables clustered. See Chapter 31, "The CLUSTER Procedure," and Chapter 100, "The VARCLUS Procedure," for details.

TYPE=UCORR Data Sets

A TYPE=UCORR data set is almost identical to a TYPE=CORR data set, except that the correlations are uncorrected for the mean. The corresponding value of the _TYPE_ variable is 'UCORR' instead of 'CORR'. Uncorrected standard deviations are in observations with _TYPE_='USTD'. A TYPE=UCORR data set can be used as input for every SAS/STAT procedure that uses a TYPE=CORR data set, except for the CANDISC, DISCRIM, and STEPDISC procedures. TYPE=UCORR data sets can be created by the CANCORR, PRINCOMP, and VARCLUS procedures.

TYPE=UCOV Data Sets

A TYPE=UCOV data set is similar to a TYPE=COV data set, except that the covariances are uncorrected for the mean. Also, the corresponding value of the _TYPE_ variable is 'UCOV' instead of 'COV'. A TYPE=UCOV data set can be used as input for every SAS/STAT procedure that uses a TYPE=COV data set, except for the CANDISC, DISCRIM, and STEPDISC procedures. TYPE=UCOV data sets can be created by the PRINCOMP procedure.

TYPE=WEIGHT Data Sets

The CALIS procedure creates and accepts as input a TYPE=WEIGHT data set. This data set contains the weight matrix used in generalized, weighted, or diagonally weighted least squares estimation. See Chapter 27, "The CALIS Procedure," for details.

Definitional Formulas

This section contrasts corrected and uncorrected SSCP, COV, and CORR matrices by showing how these matrices can be computed. In the following formulas, assume that the data consist of two variables, X and Y, with n observations.

$$\text{SSCP} = \begin{bmatrix} n & \sum X & \sum Y \\ \sum X & \sum X^2 & \sum XY \\ \sum Y & \sum XY & \sum Y^2 \end{bmatrix}$$

$$\text{CSSCP} = \begin{bmatrix} \sum(X-\bar{X})^2 & \sum(X-\bar{X})(Y-\bar{Y}) \\ \sum(X-\bar{X})(Y-\bar{Y}) & \sum(Y-\bar{Y})^2 \end{bmatrix}$$

$$\text{COV} = \frac{\text{CSSCP}}{n-1} = \frac{1}{n-1}\begin{bmatrix} \sum(X-\bar{X})^2 & \sum(X-\bar{X})(Y-\bar{Y}) \\ \sum(X-\bar{X})(Y-\bar{Y}) & \sum(Y-\bar{Y})^2 \end{bmatrix}$$

$$\text{UCOV} = \frac{1}{n}\begin{bmatrix} \sum X^2 & \sum XY \\ \sum XY & \sum Y^2 \end{bmatrix}$$

$$\text{CORR} = \begin{bmatrix} 1 & \dfrac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}} \\ \dfrac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}} & 1 \end{bmatrix}$$

$$\text{UCORR} = \begin{bmatrix} 1 & \dfrac{\sum XY}{\sqrt{\sum X^2\sum Y^2}} \\ \dfrac{\sum XY}{\sqrt{\sum X^2\sum Y^2}} & 1 \end{bmatrix}$$

Subject Index

Syntax Index

Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to **yourturn@sas.com**. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to **suggest@sas.com**.