Paper 4113-2020

%SURVEYCORRCOV Macro: Complex Survey Data Correlations for Multivariate Analysis and Model Building

David R. Nelson and Siew Wong-Jacobson, Eli Lilly & Company

ABSTRACT

SAS® SURVEY procedures cover the main topics of descriptive statistics (MEANS, FREQ) and regression (REG, LOGISTIC, PHREG). But as the use of complex surveys evolve, particularly among students who often use this data due to its high quality and low price, adding even more analytics that are suitable for this data further opens its horizons. Twelve SAS/STAT® procedures can use special SAS® data sets with the CORR and COV options as input data for analyses such as PRINCOMP, FACTOR, and VARCLUS. Having this functionality as our motivation, we extended Jessica Hampton's "PROC SURVEYCORR" approach to create a %SURVEYCORRCOV macro to include features of the CORR procedure. For example, rather than a vector of correlations, %SURVEYCORRCOV provides a matrix of correlations and their *p*-values, for both the observed values and the within-domain ranks. In addition, %SURVEYCORRCOV generates standard deviations, which can be used to create covariance matrices. The output data sets from %SURVEYCORRCOV can be used directly in procedures that use CORR and COV.

We review the parameters for using %SURVEYCORRCOV and examples for use in multivariable analyses such as principal components and factor analysis, and how variable clustering can be part of a regression modeling approach. We also provide practical advice for all data users, such as when to use the correlation or correlation matrix, and orthogonal or non-orthogonal factors.

INTRODUCTION

Complex survey data is an important source of real-world evidence, as it is often designed to be nationally representative, and processed with imbedded data quality checks. Also, complex survey data sets are often free of charge, therefore popular among students for research projects, theses, and dissertations. Unlike data collected through simple random sampling, analysis of complex survey data must take the sampling design into account, the details of which are well described elsewhere (Lewis, 2016; Heeringa et al., 2017). West et al. (2016) document that too often, publications do not properly report the correct analytic techniques for complex survey data. We recommend these three checklist steps as the foundation of analysis of complex survey data, for both unbiased estimates and proper standard errors:

1) Always use the survey procedures, e.g., SURVEYMEANS and SURVEYFREQ
2) Always use the cluster, strata, and appropriate weights (e.g., for NHANES, interview weights for questionnaires, exam weights for labs, et cetera)
3) Do not delete observations or use BY or WHERE statements. Create an analytical subset for use as a DOMAIN; analyzing the subgroup alone may create empty strata and affect the standard errors

SAS provides procedures for complex survey data univariate analysis (SURVEYMEANS and SURVEYFREQ), multivariable analysis (SURVEYREG, SURVEYLOGISTIC, and SURVEYPHREG), imputation (SURVEYIMPUTE), and the powerful tool SURVEYSELECT for a wide range of uses. However, within SAS/STAT, multivariate analysis and multivariable model selection

are not part of the SAS complex survey analyst's toolbox. Therefore, we describe a macro that expands the use of complex survey data to other SAS procedures. We extend the ideas of Hampton's (2013) "Proc SurveyCorr" approach to create a %SURVEYCORRCOV macro. Our motivation is two-fold. First is constructing a macro that mimics the key features of the CORR procedure, such as creating a matrix of pairwise correlations and *p*-values for a list of variables (including an option based on ranks) with the ability to output these to TYPE=CORR (correlation), COV (covariance), and SSCP (sum of squares and crossproducts) data sets. This provides our second motivation, using appropriate complex survey correlation and covariance matrices as input into SAS procedures such as PRINCOMP, FACTOR, VARCLUS, CANCORR, and REG, to expand the toolbox of complex survey analytics.

Initially, we describe usage and options of the %SURVEYCORRCOV macro, then we delve into its use of exporting data sets, with examples from variable selection during model building and multivariate analyses. For those analysts who do not use complex survey data, we also provide general information, for instance, when to use the correlation or covariance matrix as the input into principal components analysis. We highlight "Rules of Thumb" throughout, many of which also apply to uses of these procedures for all data types.

THE %SURVEYCORRCOV MACRO

Our approach throughout this paper is "design-based" complex survey data analysis rather than the model-based, which is consistent with the existing SURVEY procedures. The underlying backbone of the functionality of the %SURVEYCORRCOV macro is output generated by PROC SURVEYREG, and we use SURVEYREG in two ways, because this procedure can incorporate the strata, clusters, and weights of complex survey data. First, the goal is creating TYPE=CORR, COV, and SSCP data sets for use as the data source into other SAS procedures. Twelve SAS/STAT® procedures utilize CORR, COV, and SSCP matrices as input: ACECLUS, CALIS, CANDISC, DISCRIM, FACTOR, MI, MIANALYZE, PRINCOMP, REG, SIMNORM, STEPDISC, and VARCLUS. Therefore, with a correlation data set utilizing design-based characteristics expands these procedures to provide complex survey data analysis.

GENERATING COMPLEX SURVEY MATRIX DATA SETS

Generating output matrix data sets utilizes the XPX option in the model statement of PROC SURVEYREG, with steps illustrated in Figure 1. The XPX option creates a sum of squares and crossproducts matrix that will utilize the survey weights when they are included. Subsequently, the macro places this output matrix from XPX as input into PROC PRINCOMP, which in turn generates CORR and COV data sets. As shown in Figure 1, the three types of matrices are all built upon one another. Macro users can specify data set names for OUTP= (same option as PROC CORR) for the Pearson Product-Moment Correlations, OUTCOV= for the covariance matrix, and OUTSSCP= for the weighted sum of squares and crossproducts matrix. If a data set name for OUTS= is included (same option as PROC CORR for Spearman correlation coefficient data sets), variables within the domain of interest (domains are subgroups within the overall population) are ranked initially for rank-based correlations

To demonstrate uses of the %SURVEYCORRCOV macro, we use the Medical Expenditure Panel Survey (MEPS), a nationally representative survey of the U.S. civilian noninstitutionalized population. Our examples utilize the MEPS 2016 Full Year Consolidated Data File (HC-192), available at https://www.meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192. The initial demonstration features three questions from the functional status survey, the Short-Form 12 Version 2 (SF-12v2®) of Ware, et al (1996): ADCAPE42 (Felt Calm/Peaceful), ADCLIM42 (Health limits climbing stairs), and ADDAYA42 (Health

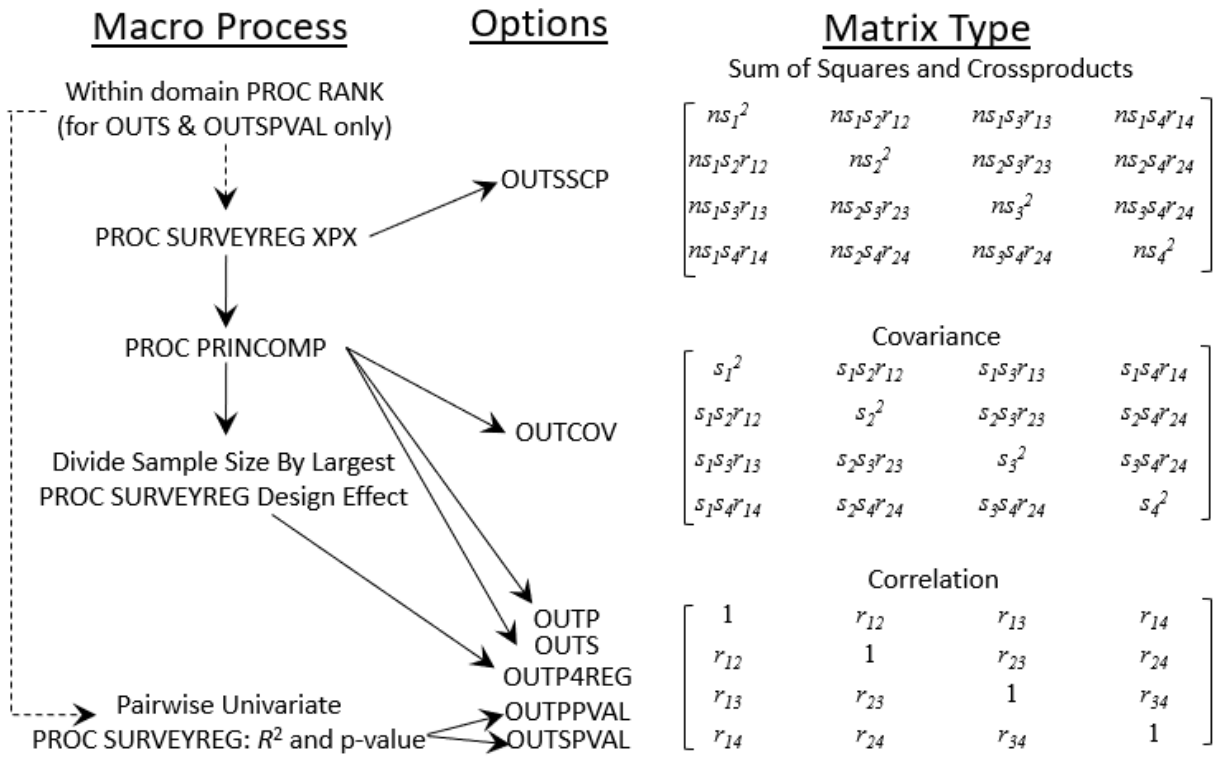


Figure 1. The %SURVEYCORRCOV macro process to derive output; Note, the use of PRINCOMP is purely to output data sets, not to perform the principal components analyses described later

OUTSSCP

	NAME	Intercept	ADCAPE42	ADCLIM42	ADDAYA42	_TYPE_
1	Intercept	168501323.71	624245608.62	459477587.78	465672726.38	SSCP
2	ADCAPE42	624245608.62	2474520974.5	1725892715.9	1746481902.6	SSCP
3	ADCLIM42	459477587.78	1725892715.9	1308699270.7	1309348330.3	SSCP
4	ADDAYA42	465672726.38	1746481902.6	1309348330.3	1336548776.8	SSCP

OUTCOV

	TYPE	_NAME_	ADCAPE42	ADCLIM42	ADDAYA42
1	MEAN		3.7046926093	2.7268485355	2.7636146478
2	N		168501323.71	168501323.71	168501323.71
3	COV	ADCAPE42	0.9607217278	0.1404706617	0.1264535757
4	COV	ADCLIM42	0.1404706617	0.3309973029	0.2345936384
5	COV	ADDAYA42	0.1264535757	0.2345936384	0.2944119861

OUTP

	TYPE	_NAME_	ADCAPE42	ADCLIM42	ADDAYA42
1	MEAN		3.7046926093	2.7268485355	2.7636146478
2	STD		0.9801641331	0.5753236505	0.5425974439
3	N		168501323.71	168501323.71	168501323.71
4	CORR	ADCAPE42	1	0.2491004971	0.2377686349
5	CORR	ADCLIM42	0.2491004971	1	0.7514952847
6	CORR	ADDAYA42	0.2377686349	0.7514952847	1

OUTS

	TYPE	_NAME_	ADCAPE42	ADCLIM42	ADDAYA42
1	MEAN		8256.6705105	8576.0056083	8642.6868035
2	STD		4317.0545466	3410.5458954	3242.7698343
3	N		168501323.71	168501323.71	168501323.71
4	CORR	ADCAPE42	1	0.2505365627	0.2411067811
5	CORR	ADCLIM42	0.2505365627	1	0.7130382043
6	CORR	ADDAYA42	0.2411067811	0.7130382043	1

OUTP4REG

	TYPE	_NAME_	LogTotalExp	ADCAPE42	ADCLIM42	ADDAYA42
1	MEAN		6.0827598657	3.7046926093	2.7268485355	2.7636146478
2	STD		3.159605705	0.9801641331	0.5753236505	0.5425974439
3	CORR	LogTotalExp	1	-0.148583746	-0.248655341	-0.236652364
4	CORR	ADCAPE42	-0.148583746	1	0.2491004971	0.2377686349
5	CORR	ADCLIM42	-0.248655341	0.2491004971	1	0.7514952847
6	CORR	ADDAYA42	-0.236652364	0.2377686349	0.7514952847	1
7	N		.	7085.2344673	16394.192002	12115.135918

Display 1. Five of the options for output data sets from %SURVEYCORRCOV

limits moderate activities), within the domain of the age subgroup 20 to 65. Display 1 provides the results for datasets created as correlation, rank-based correlation, covariance, and SSCP matrices. Notice for these data sets, the "N" represents the sum of the weights, in this case, over 168 million observations. For many of the multivariate procedures such as principal components or factor analysis, the "N" is not utilized, only the correlation or covariance matrix, therefore the sample size does not affect the results. However, to gain the functionality of PROC REG, for example, requires a meaningful "N", and using the upper four data sets in Display 1 in REG provides standard errors of regression estimates based on over hundred million observations, hence that are too small, and resulting in smaller *p*-values.

We include another data set option, OUTP4REG, for procedures that require reasonable sample size estimates, such as the use of PROC REG that we discuss in more detail later in the multivariable regression section. We illustrate here the creation of OUTP4REG with three SF-12 variables from above as independent variables, and log of total medical expenses during 2016 as the dependent variable in Table 1. Using the OUTP matrix with a N of greater than 168 million in PROC REG produces the same regression coefficients as PROC SURVEYREG, but much smaller standard errors, and hence, higher *t* and smaller *p*-values (Table 1; A versus B). %SURVEYCORRCOV divides the survey's observed sample size *n* by the "Design Effect" (output by SURVEYREG; Table 1A) for each variable, and the resulting "N" in OUTP4REG is a sample size that produces the results close to PROC SURVEYREG. The design effect quantifies the ratio of the observed variance to the variance computed under the assumption of simple random sampling. Note, in the bottom OUTP4REG data set in Display 1, the N's all vary. %SURVEYCORRCOV divides all the variables by their own design effect, however, PROC REG uses the *n* for analysis based on the smallest N in the matrix, hence, the largest design effect. Note in Table 1, the independent variable ADCAPE42 has the largest design effect, hence this is used for the sample size in the REG procedure. The results in Table 1, A versus C, are identical for ADCAPE42, but the standard errors for the other variables are larger using PROC REG. Therefore, the regression analysis using OUTP4REG and PROC REG is more "conservative" in statistical power than PROC SURVEYREG. We believe that for the uses of PROC REG that we describe later, that this conservative approach is still very beneficial to utilize REG's capabilities.

	A) SURVEYREG results					B) REG results with OUTP data				C) REG results with OUTP4REG data			
Parameter	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Design Effect**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	11.024	0.157	70.00	<.0001	1.28	11.024	0.001	7941.50	<.0001	11.024	0.214	51.48	<.0001
ADCAPE42	**-0.280**	**0.038**	**-7.33**	**<.0001**	**2.36**	-0.280	*0.000*	*-1130.10*	<.0001	**-0.280**	**0.038**	**-7.33**	**<.0001**
ADCLIM42	-0.816	0.063	-12.97	<.0001	1.02	-0.816	0.001	-1315.30	<.0001	-0.816	0.096	-8.53	<.0001
ADDAYA42	-0.607	0.077	-7.85	<.0001	1.38	-0.607	0.001	-925.68	<.0001	-0.607	0.101	-6.00	<.0001

Table 1. Comparison of results from SURVEYREG versus OUTP and OUTP4REG data sets from %SURVEYCORRCOV. Note ADCAPE42 results are same for A and C.

As a side note, the data sets OUTP and OUTP4REG provide standard deviations, which are not provided by SAS SURVEY procedures. Therefore, these two data sets can be used or transposed to produce a table of means and standard deviations.

CREATING PROC CORR-LIKE OUTPUT WITH SIGNIFICANCE TESTS

The second use of PROC SURVEYREG is creating output typical of PROC CORR, such as producing a matrix of correlations and their associated pairwise *p*-values. Previously, SAS code to produce equivalents of Pearson product-moment correlations was proposed by Hampton (2013), for a Proc SurveyCorr capability. The idea behind Hampton's code is performing a series of univariate PROC SURVEYREG analyses, and then taking the square

root of the resulting R^2 values as correlations, and the *F*-test *p*-value as the significance test. Hampton's code takes a single variable as a dependent variable, cycles through a list of *j* independent variables to perform *j* PROC SURVEYREG runs, and creates a *j* by 1 vector of correlations and accompanying *p*-values. We extend this approach in the %SURVEYCORRCOV macro (Figure 1, lower part).

The %SURVEYCORRCOV macro takes the user's list of *j* variables and constructs a matrix of correlations with *p*-values by performing $j^2 - j$ PROC SURVEYREG analyses with each pairwise combination of variables. Because correlation matrices are symmetric, it may seem redundant to perform both "model y=x" and "model x=y." However, with complex survey data, the *p*-values may differ between these two model statements. Therefore, the macro produces both *p*-values.

In general, our experience is that the two *p*-values resulting from "model y=x" and "model x=y" are usually very close. If one prefers being conservative with type I error for hypothesis testing, use the greater of the two *p*-values. If one wants to maintain the uniform distribution of *p*-values under the null hypothesis, for instance for use in PROC MULTTEST, randomizing the order of the variables for entry into the parameter VARLIST list, and then randomly selecting the lower left or upper right half of the correlation matrix should suffice.

Again, %SURVEYCORRCOV provides correlations and *p*-values based on ranks of observations within the subgroup/domain of interest. After PROC RANK creates within domain ranks, then the same process of pairwise SURVEYREG analyses produces rank-based correlations. Please note, the use of ranks and Spearman correlations have not been reported in the literature for complex survey data. Inclusion of rank-based correlation matrix in our macro is mainly as output for use as a data set for procedures such as PROC VARCLUS, as we describe later. Though rank-based correlations are provided with *p*-values, please regard them as useful but not for hypothesis testing at this time.

Display 2 illustrates the output of the Pearson correlations and *p*-values for three SF-12 variables using both Excel and Listing options. If one needs only the correlation matrices without *p*-values, it is best to specify OUTP or OUTP4REG rather than OUTPPVAL, because of computational speed. OUTPPVAL and OUTSPVAL utilize more processing time because they require pairwise SURVEYREGs, whereas OUTP and OUTP4REG utilize a single SURVEYREG run.

Pearson Correlation Matrix

Rank-ordered data is not used

parameter	adcape42	adclim42	addaya42
adcape42	1	0.249	0.2376
		<.0001	<.0001
adclim42	0.249	1	0.7519
	<.0001		<.0001
addaya42	0.2376	0.7519	1
	<.0001	<.0001	

Pearson Correlation Matrix

Rank-ordered data is not used

parameter	adcape42	adclim42	addaya42
adcape42	1.0000	0.2490	0.2376
		<.0001	<.0001
adclim42	0.2490	1.0000	0.7519
	<.0001		<.0001
addaya42	0.2376	0.7519	1.0000
	<.0001	<.0001	

Display 2. Two output options (Excel and Listing) similar to PROC CORR for the within domain correlation matrix and *p*-values.

%SURVEYCORRCOV MACRO PARAMETERS

The parameters for the SAS macro %SURVEYCORRCOV are in Table 2.

DATA=	(Required) SAS data set name
STRATA=	For all SURVEY procs, the STRATA statement names variables that form the strata in a stratified sample design
CLUSTER=	For all SURVEY procs, the CLUSTER statement names variables that identify the clusters in a clustered sample design
WEIGHT=	For all SURVEY procs, the WEIGHT statement names the variable that contains the sampling weights
DOMAIN=	(Required) In SAS SURVEY procs, the DOMAIN statement names the variables that identify subgroups of interest, or "domains". For this macro, the variable that identifies a subpopulation of interest for calculation of the correlation, covariance, and/or SSCP. For the entire sample to be included in the macro, create a variable and assign the same value, and use this variable as DOMAIN
SUBGRP=	(Required) The level of the DOMAIN statement to be included for analysis. For instance, if the subgroup of analysis is defined as include=1, then specify DOMAIN=include, and SUBGRP=1. Do not use value -999 if NOMISS=1
VARLIST=	(Required unless only output is OUTP4REG) The variables comprising the CORR/COV/SSCP matrix, separated by spaces
EXCELOUT=	if EXCELOUT=1, then create excel output file with a tab for each data set specified in any of the OUT* parameters
EXCELFILE=	Path and filename for Excel file if EXCELOUT=1
SURVEYREGOPTIONS=	Specify options as the would appear in the PROC SURVEYREG statement, such as missing value handling and variance estimation methods. For example, "SURVEYREGOPTIONS = nomcar varmethod=jackknife" specifies treating missing values as not missing completely at random, and Jackknife as the variable estimation method. If specified, all analyses will use these options
Parameters for Correlations with *p*-values	
ID=	(Required for OUTPPVAL and OUTSPVAL) Identification variable, unique for each observation
NVAR=	(Required for OUTPPVAL and OUTSPVAL) Integer representing the number of variables listed in VARLIST
OUTPPVAL=	Data set name for output data set with Pearson Product-Moment Correlation and their *p*-values
OUTSPVAL=	Data set name for output data set with within-domain rank-based correlation (Spearman's correlation) and their *p*-values
NOMISS=	For *p*-value matrices, similar option as PROC CORR. NOMISS=1 excludes observations with any VARLIST missing values from the analysis. Otherwise the correlations and *p*-values are from all available data. NOTE, for all output data sets without *p*-values, NOMISS is used

LISTING=	If LISTING=1, then PROC PRINT will display OUTPPVAL and OUTSPVAL, if specified
Parameters for Correlations without *p*-values	
OUTP=	Similar to PROC CORR option; Data set name for output data set with Pearson Product-Moment Correlation, no *p*-values
OUTS=	Similar to PROC CORR; Data set name for output data set with within-domain rank-based correlation, no *p*-values
OUTCOV=	Data set name for output data set with covariance matrix
OUTSSCP=	Data set name for output data set with sum of squares and cross-products matrix
OUTP4REG=	Data set name for output data set with Pearson Product-Moment Correlation, no *p*-values. N is adjusted by the design effect for each variable
DEPEND=	(Required for OUTP4REG) Dependent variable name that corresponds to SURVEYREG model statement before the "="
INDEP=	(Required for OUTP4REG) Independent variable names that correspond to SURVEYREG model statement after the "="

Table 2. Parameters for the SAS macro %SURVEYCORRCOV

TOOLS FOR COMPLEX SURVEY DATA MULTIVARIABLE REGRESSION

Additional multivariable regression tools can be accessed with data sets from %SURVEYCORRCOV. Multivariable regression model building is an extensive and sometimes controversial concept. Full discussions of regression model building strategies (Harrell, 2015) and comparison of methods when collinearity is present (Dormann et al., 2013) are well discussed elsewhere, and great care needs to be taken when performing these analyses. For SAS users, Wang and Shin (2011) provide macros %StepSvylog, that use PROC SURVEYLOGISTIC, and %StepSvyreg that utilize PROC SURVEYREG, for forward, backward, and stepwise selection, but we believe that %SURVEYCORRCOV further extends the capabilities of multivariable regression in SAS.

VARIABLE CLUSTERING USING PROC VARCLUS

In an age of increasingly more variables available to analysts, both variable reduction and dealing with collinearity are common needs. One extremely informative tool is variable clustering, as implemented with PROC VARCLUS. Nelson (2001) explains uses of VARCLUS in detail. The procedure starts with all variables, and then continues splitting the variables into correlated groups until a stopping rule is achieved. The final results are distinct sets of highly correlated (positive or negative) variables.

Using variable clustering in multivariable model building is based on a different philosophy than stepwise, backward, and forward selection. One variable clustering approach selects variables to represent the entire independent variable or covariate space, rather than looking for variable significance. One tactic for this approach is: 1) take all of the potential independent variables and create a Spearman rank-based correlation matrix; 2) use variable clustering until either the procedure stops splitting clusters or the number of clusters equals the sample size (for linear regression) or number of events (for logistic regression) divided by 20 (therefore, at least 20 observations/events per variable in the model); 3) for independent variables, use a score of all of the variables within a cluster, or take a representative variable from the cluster (i.e., least missing values, most

representative of the cluster, et cetera). 4) Use these summary scores or representatives of the clusters as independent variables, and then include them in the model regardless of statistical significance.

An illustration of the process will include the twelve SF-12 questions, nine types of log-transformed medical expenditures (inpatient, outpatient, office-based, emergency, medications, supplies, home health, dental, and vision), and age from MEPS 2016, as 22 hypothetical independent variables. The medical expenditures are paid by all sources throughout the year for each person in the survey. Therefore, throughout this paper, discussion of health care costs is not "out of pocket," but all sources of payment including insurance payments plus out of pocket. %SURVEYCORRCOV creates a rank-based correlation of these 22 variables, and we then run PROC VARCLUS with the OUTS data set:

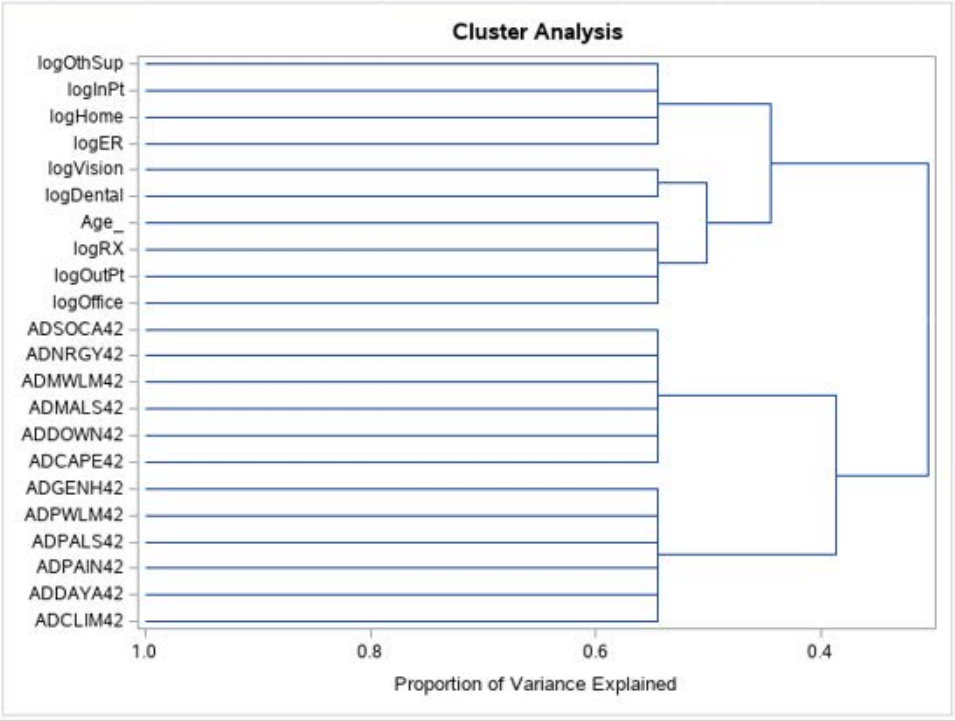
```
title1 "Using the %SURVEYCORRCOV matrix to produce the";
title2 "rank-based correlation matrix of nine types of expenses";
title3 "twelve SF-12 questions, age from MEPS 2016 data";
title4 "Note: Subsequent %SurveyCorrCov calls will only include";
title5 "macro variables with assigned values";
%SurveyCorrCov(data=a, id=, nvar=, strata=VARSTR, cluster=VARPSU,
   weight=PERWT16F, domain=include, subGrp=1, outp=, outs=RankCorr,
   outsscp=,outcov=,depend=,indep=,outp4reg=,surveyregoptions=,
   outppval=, outspval=, nomiss=, excelout=, excelfile=, listing=,
   varlist=ADCAPE42 ADCLIM42 ADDAYA42 ADDOWN42 ADMALS42 ADMWLM42
   ADPAIN42 ADPALS42 ADPWLM42 ADNRGY42 ADSOCA42 logDental logER
   logHome logOffice logOutPt logRX logInPt logOthSup logVision
   age16X ADGENH42);

title1 "Variable Clustering";
   proc varclus data=RankCorr(TYPE=CORR);
        var ADCAPE42 ADCLIM42 ADDAYA42 ADDOWN42
             ADMALS42 ADMWLM42 ADPAIN42 ADPALS42 ADPWLM42
             ADNRGY42 ADSOCA42 logDental logER logHome logOffice
             logOutPt logRX logInPt logOthSup logVision age16X
             ADGENH42;
```

Notice during the macro call, we use the clusters, strata, and weights from the data set from MEPS. In VARCLUS, we use the TYPE=CORR data that includes these complex survey design attributes rather than the original, by individual MEPS data set, because VARCLUS does not have options for these design elements like the SURVEYREG procedures. The domain is the age group 20-65, from a variable named "include."

With these hypothetical independent variables, variable clustering indicates that there are five correlated clusters of variables (Display 3). Cluster 1 is comprised of six SF-12 questions about general health, pain, climbing stairs, et cetera. Cluster 2 includes age and three types of medical expenditures (office-based, outpatient, and medications). Cluster 3 are the remaining six SF-12 question, focusing on energy and mental health, whereas Cluster 4 contains expenditures from ER, Home Health, Inpatient, and Supplies. Finally, Cluster 5 contains both the Dental and Vision expenditures. Therefore, to create a model, the strategy is to use five independent variables. They can be summarized using the Standardized Scoring Coefficients (Display 3) creating a weighted sum for each cluster. Another strategy is to use the variable most representative of the cluster, and this can be achieved by using the smallest 1-R^2 ratio value within a cluster (Display 3). In this example, the five clusters were far below the cutoff of "number of observations divided by 20," but for guidance on the number of observations to use in this case, use OUTP4REG's use of the design effect to estimate an operational sample size.

5 Clusters		R-squared with		
Cluster	Variable	Own Cluster	Next Closest	1-R**2 Ratio
Cluster 1	ADCLIM42	0.6588	0.2134	0.4337
	ADDAYA42	0.6736	0.2057	0.4109
	ADPAIN42	0.5867	0.2645	0.5619
	ADPALS42	0.7178	0.3368	0.4256
	ADPWLM42	0.7871	0.3260	0.3159
	ADGENH42	0.4196	0.2519	0.7757
Cluster 2	logOffice	0.6555	0.0920	0.3794
	logOutPt	0.3283	0.0424	0.7015
	logRX	0.6935	0.1539	0.3623
	Age_	0.3236	0.0836	0.7381
Cluster 3	ADCAPE42	0.4952	0.1281	0.5790
	ADDOWN42	0.6088	0.1629	0.4673
	ADMALS42	0.6785	0.2892	0.4523
	ADMWLM42	0.5862	0.2984	0.5897
	ADNRGY42	0.4922	0.2609	0.6870
	ADSOCA42	0.6443	0.3300	0.5309
Cluster 4	logER	0.4521	0.0455	0.5740
	logHome	0.2880	0.0338	0.7369
	logInPt	0.4977	0.0396	0.5230
	logOthSup	0.2659	0.0428	0.7669
Cluster 5	logDental	0.5612	0.0657	0.4697
	logVision	0.5612	0.0202	0.4478

Standardized Scoring Coefficients					
Cluster	1	2	3	4	5
ADCAPE42	0.000000	0.000000	0.200754	0.000000	0.000000
ADCLIM42	0.211176	0.000000	0.000000	0.000000	0.000000
ADDAYA42	0.213529	0.000000	0.000000	0.000000	0.000000
ADDOWN42	0.000000	0.000000	0.222598	0.000000	0.000000
ADMALS42	0.000000	0.000000	0.234995	0.000000	0.000000
ADMWLM42	0.000000	0.000000	0.218431	0.000000	0.000000
ADPAIN42	0.199287	0.000000	0.000000	0.000000	0.000000
ADPALS42	0.220417	0.000000	0.000000	0.000000	0.000000
ADPWLM42	0.230816	0.000000	0.000000	0.000000	0.000000
ADNRGY42	0.000000	0.000000	0.200145	0.000000	0.000000
ADSOCA42	0.000000	0.000000	0.228991	0.000000	0.000000
logDental	0.000000	0.000000	0.000000	0.000000	0.667443
logER	0.000000	0.000000	0.000000	0.447160	0.000000
logHome	0.000000	0.000000	0.000000	0.356899	0.000000
logOffice	0.000000	0.404646	0.000000	0.000000	0.000000
logOutPt	0.000000	0.286349	0.000000	0.000000	0.000000
logRX	0.000000	0.416198	0.000000	0.000000	0.000000
logInPt	0.000000	0.000000	0.000000	0.469193	0.000000
logOthSup	0.000000	0.000000	0.000000	0.342910	0.000000
logVision	0.000000	0.000000	0.000000	0.000000	0.667443
Age_	0.000000	0.284301	0.000000	0.000000	0.000000
ADGENH42	0.168535	0.000000	0.000000	0.000000	0.000000



Display 3. PROC VARCLUS output for twelve SF-12 questions, nine types of log-transformed medical expenditures, and age from MEPS 2016, based on rank-based correlation from %SURVEYCORRCOV

USE OF PROC REG WITH OUTPUT FROM %SURVEYCORRCOV

With the availability of PROC SURVEYREG, why is PROC REG useful when analyzing complex survey data? Although SURVEYREG properly utilizes the strata, clusters, and weights of the survey design for multivariable regression, it lacks some of REG's utilities. In this section, we examine two: variance inflation factors/ridge regression and variable selection. Our

example will use the logarithm of total medical care expenses as the dependent variable, and age and the SF-12 questions as the independent variables:

```
title1 "Using the %SURVEYCORRCOV matrix to produce the";
title2 "correlation matrix for dependent variable of";
title3 "log of total medical expenditures in 2016,";
title4 "with independent variables of twelve SF-12";
title5 "and age from MEPS data";
%SurveyCorrCov(data=a, strata=VARSTR, cluster=VARPSU,
        weight=PERWT16F, domain=include, subGrp=1,
             outp4reg=ProcReg, depend=LogTotalExp,
        indep=ADCAPE42 ADCLIM42 ADDAYA42 ADDOWN42
             ADMALS42 ADMWLM42 ADPAIN42 ADPALS42 ADPWLM42
             ADNRGY42 ADSOCA42 ADGENH42 age16X);

ods graphics on;
title1 "Variance Inflation Factor and Ridge Regression";
proc reg data=ProcReg(TYPE=CORR)  outvif
     outest=b ridge=0 to 2 by .2;
   model LogTotalExp = ADCAPE42 ADCLIM42 ADDAYA42 ADDOWN42
        ADMALS42 ADMWLM42 ADPAIN42 ADPALS42 ADPWLM42
        ADNRGY42 ADSOCA42 ADGENH42 age16X /vif;

proc print data=b;

title1 "Model Selection By Mallows' CP Selection";
proc reg data=ProcReg(TYPE=CORR)  ;
   model LogTotalExp = ADCAPE42 ADCLIM42 ADDAYA42 ADDOWN42
        ADMALS42 ADMWLM42 ADPAIN42 ADPALS42 ADPWLM42
        ADNRGY42 ADSOCA42 ADGENH42 age16X / selection=cp;

title1 "Implementing Mallows' CP Model from PROC REG in SURVEYREG";
PROC SURVEYREG DATA=a;
        MODEL LogTotalExp= ADCLIM42 ADDOWN42 ADMALS42 ADMWLM42
             ADPAIN42 ADPWLM42 ADNRGY42 Age16X/deff;
        strata VARSTR;
        cluster VARPSU;
        weight PERWT16F;
        domain include; run;
```

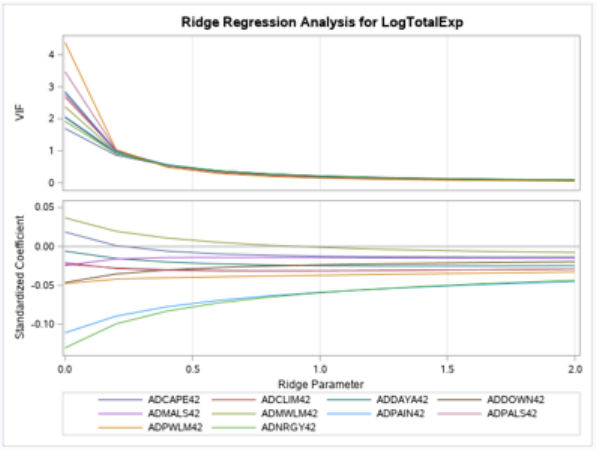
As noted previously, the use of ever-increasing numbers of potential independent variables is expanding in the real-world evidence space in general, including complex survey data. One tool to assess the severity of multicollinearity in a multivariable model is the “Variance Inflation Factor” (VIF). VIFs for our MEPS regression example are included in Display 4. Cutoffs of VIFs, such as being greater than four or ten, are used in practice to identify collinearity at a level that must be accounted for. Interpret VIF as the square of the increase in the standard error compared to when all of the independent variables are uncorrelated. For instance, the only independent variable in Display 4 with a VIF exceeding “4” is ADPWLM42 (work limitations because of physical problems). A value of greater than “4” means that the standard error of the regression coefficient of ADPWLM42 is at least two (the square of “4”) times larger than it would be if it was uncorrelated with all of the other independent variables.

One method to perform regression in the face of multicollinearity is "Ridge Regression." Ridge regression adds to the diagonals of the correlation matrix, which would normally be "1", a small bias or a k-value (this amount added the diagonal value to "1" in the correlation matrix is the "ridge"). The traditional least squares approach is unbiased but struggles with collinearity due to variance inflation. The goal of ridge regression is to add just enough bias to make the regression estimates reasonably reliable approximations of the true population values.

Display 4 demonstrates the implementation of Ridge Regression. Similar to Least Absolute Shrinkage and Selection Operator (LASSO) regression, the figure in Display 4 illustrates the regression estimates changing as one goes from least squares (k=0) to increasing ridge (k>0) additions. This plot assists understanding what value of the ridge parameter results in smallest value of k after the regression coefficients stabilize. In our example, the ridge parameters from 0.25 to 1 result in low VIFs and stabilized regression coefficients, and the regression coefficients that correspond to these k values are in the data set "b" (not shown).

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	8.80293	0.27002	32.60	<.0001	0
ADCAPE42	1	0.06016	0.04403	1.37	0.1719	1.69455
ADCLIM42	1	-0.12572	0.09440	-1.33	0.1829	2.68182
ADDAYA42	1	-0.03564	0.10368	-0.34	0.7311	2.84823
ADDOWN42	1	-0.15729	0.05053	-3.11	0.0019	2.04325
ADMALS42	1	-0.08399	0.05970	-1.41	0.1595	2.77024
ADMWLM42	1	0.13978	0.06114	2.29	0.0223	2.37523
ADPAIN42	1	-0.34191	0.04608	-7.42	<.0001	2.06503
ADPALS42	1	-0.06311	0.05945	-1.06	0.2885	3.46759
ADPWLM42	1	-0.15074	0.06882	-2.19	0.0285	4.38604
ADNRGY42	1	-0.41717	0.04615	-9.04	<.0001	1.91893
ADSOCA42	1	0.01638	0.05323	0.31	0.7583	2.37368
ADGENH42	1	-0.05419	0.04292	-1.26	0.2068	1.62107
Age_	1	0.04659	0.00262	17.81	<.0001	1.10972



Display 4. PROC REG use of correlation created by %SURVEYCORRCOV, including Variance Inflation Factors and Ridge Regression results

Another value of using PROC REG is utilizing various model selection methods, none of which are part of PROC SURVEYREG. For instance, the next example uses Mallows' C_p measure to assess models, as implemented by "SELECTION=CP" in PROC REG's MODEL statement. Optimal values of Mallows' C_p are those near the value of the number of independent variables plus one (or equivalently, number of independent variables including the intercept). In the results in the Display 6, the model with eight variables has a C_p of 8.98, very close to the target, indicating a precise model with unbiased regression coefficients and successfully predicting future responses. A C_p >9 would indicate that the regression model is over-fitted (too many independent variables and possible collinearity), whereas C_p<8 may indicate a regression model is underspecified (i.e., at least one important independent variable is not included).

We include these eight independent variables in SURVEYREG as a follow-up. In all cases, the use of PROC REG with %SURVEYCORRCOV data sets is designed to be informative, but the final proper analysis is SURVEYREG. The OUTP4REG data set will only have an estimated sample size based on design effects rather than the correct SURVEYREG analysis. Note in Display 5, the PROC REG estimate of this model's R^2 is 14.65% whereas the SURVEYREG result for the same model is R^2=14.66, almost identical, and indicative that this transition from model building in PROC REG to quantification in PROC SURVEYREG resulted in a very close R^2.

PROC REG results for "selection=cp"

Number in Model	C(p)	R-Square	Variables in Model
8	8.9840	0.1465	ADCLIM42 ADDOWN42 ADMALS42 ADMWLM42 ADPAIN42 ADPWLM42 ADNRGY42 Age_
7	9.1029	0.1463	ADCLIM42 ADDOWN42 ADMWLM42 ADPAIN42 ADPWLM42 ADNRGY42 Age_
9	9.1131	0.1467	ADCAPE42 ADCLIM42 ADDOWN42 ADMALS42 ADMWLM42 ADPAIN42 ADPWLM42 ADNRGY42 Age_
8	9.2464	0.1465	ADCLIM42 ADDOWN42 ADMWLM42 ADPAIN42 ADPWLM42 ADNRGY42 ADGENH42 Age_
9	9.2598	0.1467	ADCLIM42 ADDOWN42 ADMALS42 ADMWLM42 ADPAIN42 ADPWLM42 ADNRGY42 ADGENH42 Age_
10	9.3339	0.1469	ADCAPE42 ADCLIM42 ADDOWN42 ADMALS42 ADMWLM42 ADPAIN42 ADPWLM42 ADNRGY42 ADGENH42 Age_
8	9.3644	0.1465	ADCLIM42 ADDOWN42 ADMWLM42 ADPAIN42 ADPALS42 ADPWLM42 ADNRGY42 Age_
8	9.5264	0.1464	ADCAPE42 ADCLIM42 ADDOWN42 ADMWLM42 ADPAIN42 ADPWLM42 ADNRGY42 Age_
9	9.6086	0.1466	ADCAPE42 ADCLIM42 ADDOWN42 ADMWLM42 ADPAIN42 ADPWLM42 ADNRGY42 ADGENH42 Age_
9	9.6579	0.1466	ADCLIM42 ADDOWN42 ADMALS42 ADMWLM42 ADPAIN42 ADPALS42 ADPWLM42 ADNRGY42 Age_

PROC SURVEYREG results for model with lowest C(p) above

Fit Statistics	
R-Square	0.1466
Root MSE	2.9197
Denominator DF	203

Estimated Regression Coefficients					
Parameter	Estimate	Standard Error	t Value	Pr > |t|	Design Effect
Intercept	8.7763145	0.22016506	39.86	<.0001	1.49
ADCLIM42	-0.1539057	0.06055559	-2.54	0.0118	1.14
ADDOWN42	-0.1408814	0.03939241	-3.58	0.0004	1.59
ADMALS42	-0.0838434	0.04441298	-1.89	0.0605	1.22
ADMWLM42	0.1322754	0.04439891	2.98	0.0032	1.13
ADPAIN42	-0.3509453	0.03402768	-10.31	<.0001	1.20
ADPWLM42	-0.2082989	0.04470978	-4.66	<.0001	1.43
ADNRGY42	-0.4013059	0.03590845	-11.18	<.0001	1.68
Age_	0.0471091	0.00247579	19.03	<.0001	1.90

Display 5. Top ten results from PROC REG using Mallows' C_p selection method and applying the resulting model to PROC SURVEYREG.

MULTIVARIATE ANALYSIS FOR COMPLEX SURVEY DATA

In addition to utilizing the tools for multivariable regression with complex survey data, %SURVEYCORRCOV data sets also allow access to multivariate procedures. In essence, to be able to perform PROC SURVEYPRINCOMP, PROC SURVEYFACTOR, et cetera, analyses. Many multivariate analyses start in earnest by processing the correlation or covariance matrix, and not the observation level data. Therefore, if %SURVEYCORRCOV creates these matrices while utilizing the characteristics of the complex survey design, survey analysts gain this capability in SAS.

PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis, or PCA, is a valuable tool for visualizing and understanding both the relationships among variables and the sources of total variability. There are no dependent variables in PCA, so it is classified as a tool for "unsupervised learning." The functional purpose of PCA is to create uncorrelated linear combinations of variables, and these are ordered starting from Principal Component I, which accounts for the most variability, and then sequentially to the Principal Component equal to the number of original variables. Often, one looks to determine whether the first two or three Principal Components account for a high percentage of the total variability. Therefore, determining whether fewer dimensions represent the variables of interest, hence a "dimension reduction" tool.

To illustrate PCA, we again use an example from the 2016 full-year MEPS data, and utilize nine types of medical expenditures paid by all sources throughout the year for each person in the survey. We compare three approaches to PCA with these nine variables: using the correlation matrix, the covariance matrix, and the covariance matrix of the log-transformed variables. We implement the macro and procedures as follows:

```
   title1 "Using the %SURVEYCORRCOV matrix to produce the";
   title2 "correlation matrix of nine types of expenses";
   title3 "from MEPS 2016 data. Both original and natural";
   title4 "log-tranformed variable are included";

   %SurveyCorrCov(data=a, strata=VARSTR, cluster=VARPSU,
           weight=PERWT16F, domain=include, subGrp=1, outp=MEPScorr,
           varlist= DVTEXP16 ERTEXP16 HHEXP16 OBVEXP16 OPTEXP16
                RXEXP16 IPTEXP16 OTHEXP16 VISEXP16
                logDental logER logHome logOffice logOutPt logRX
                logInPt logOthSup logVision);

   title1 "PCA based on the correlation matrix";
   proc princomp data=MEPScorr(TYPE=CORR);
           var DVTEXP16 ERTEXP16 HHEXP16 OBVEXP16 OPTEXP16 RXEXP16
                IPTEXP16 OTHEXP16 VISEXP16;

   title1 "PCA based on the covariance matrix";
   title1 "COV can be specified with CORR, COV, or SCCP input";   proc
princomp data=MEPScorr(TYPE=CORR) COV;
           var DVTEXP16 ERTEXP16 HHEXP16 OBVEXP16 OPTEXP16 RXEXP16
                IPTEXP16 OTHEXP16 VISEXP16;

   title1 "PCA based on the cov matrix of natural log-transformed";
   proc princomp data=MEPScorr(TYPE=CORR) COV;
           var logDental logER logHome logOffice logOutPt logRX
                logInPt logOthSup logVision; run;
```

Even though we perform analyses based on the covariance matrix, we only request the correlation matrix with OUTP=, because PRINCOMP will generate the covariance matrix, as it is a combination of the correlations and standard deviations in the TYPE=CORR data set, as illustrated in Figure 1. Also, even though we do not analyze the raw and log-transformed data together, we include both sets of variables in the macro, and specify which variables are needed in the PROC's VAR statement.

Results for the three approaches to PCA are included in Table 3. The first two eigenvectors for each type analysis are shown, and these are a linear combination of variables; similar to regression coefficients. But instead of predicting an outcome, these PCA coefficients account for the most variability of the nine correlated variables. Notice something that is exceedingly common, whether it is the SF-12 questions, lab measures, or body measurements: all of the first Principal Component (PC I) values are positive, no matter which matrix is used. This common result is due to all of the variables being positively correlated. Though to different degrees, as one expense goes up, the others tend to also increase. These positive coefficients, so often observed in PC I, allow for interpretations similar to that of morphometrics, the multivariate analysis of size and shape. PC I describes the overall "size" of medical expenditures, or a weighted sum. All other Principal Components, from PC II to PC IX, in this example, will be uncorrelated with PC I, and hence can be thought of accounting for the typical "shapes" of medical expenditure use seen in the US population.

But which of these three approaches are appropriate? Initially, the PROC PRINCOMP's default, the correlation matrix is utilized. When should this matrix be used for PCA? Using the correlation matrix has one interesting attribute: it provides PCA on the variables but treats them all as standardized, as if all their means are "0" and standard deviations are "1"; no PROC STANDARD required. This can be seen in the differences between the

correlation and covariance illustrated in Figure 1. The effect of using standardized variables is beneficial when one includes a mixture of variables of different scales in the PCA (e.g., blood pressure and cholesterol measures) or the variability is constructed by the designer in a patient reported scale, such as the SF-12. The variability of answers to each question is partially based on how the scale developer derives the number of levels, so standardizing these variables is reasonable. Therefore, PCA based on correlation matrices is extremely useful when the variables are heterogenous in scale and units.

However, our MEPS expenditure example includes variables with the same units, dollars. For expenditures, analyses of body measurements, and other types of measures, accounting for the actual variability is meaningful rather than standardizing all variables. Therefore, the covariance matrix provides a PCA solution based on the actual variability of the measures. The issue with this approach is shown in the results in Table 3 for the covariance matrix. Almost all of the first PC loadings are based on inpatient stay, with a PC I loading of 0.961, and the other variables have minimal values. This is due to the high variability among inpatient expenditures compared to other types. Within our age-based domain, most individuals have $0 for inpatient stays (94.2%), whereas a subset have over $100,000 of inpatient expenditures (5.7%), with only 0.1% between those extremes, creating far more variability than the other cost categories. In some cases, this may be desirable, since this PC I accounts for most of the variability. However, there is an intermediate between standardizing all variables by using the correlation matrix and the dominance of highly variable measures in the covariance approach.

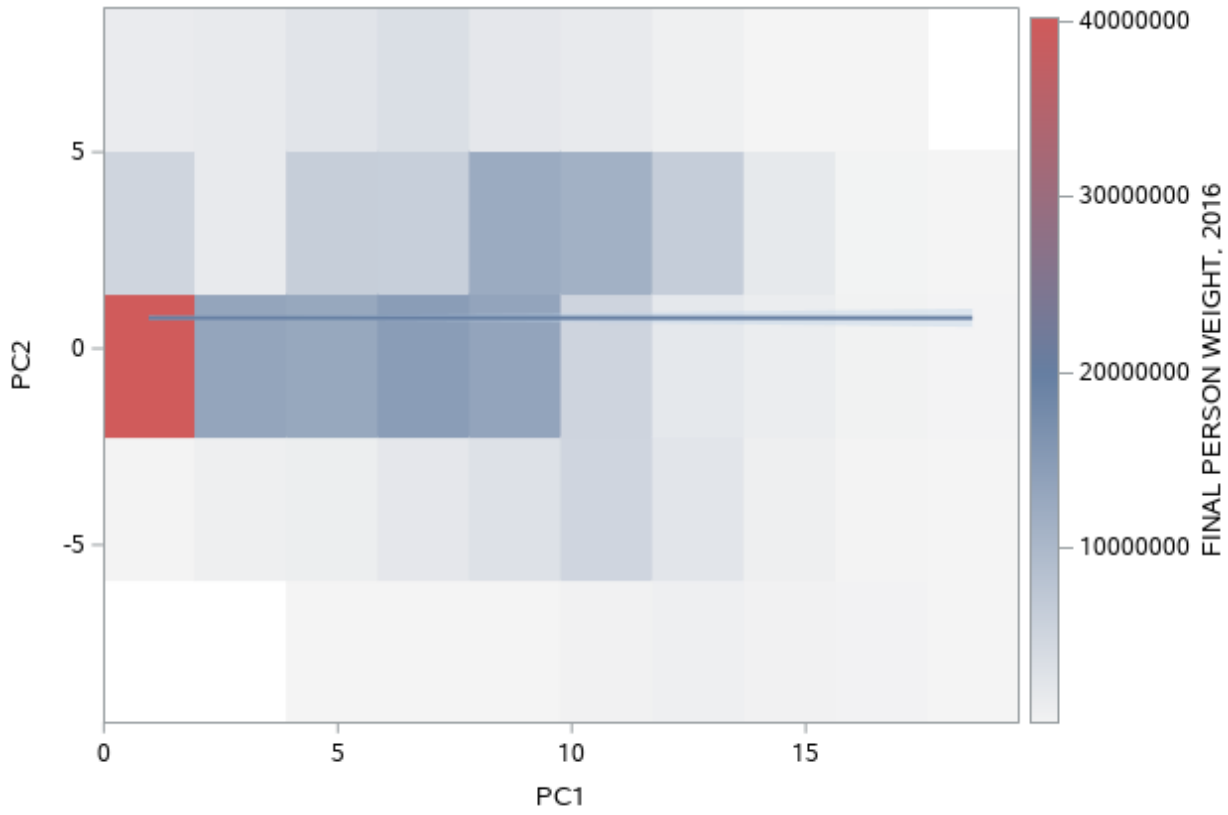
This compromise utilizes the logarithm of costs in the macro for correlation/covariance matrix creation. Taking the natural logarithm (the SAS LOG function) after the addition of the value of "1" to each variable if there are zeroes in the data (the addition of LOG[Value+1] results in "0" if the initial value is zero) is a typical strategy. By taking the logarithm, the PCA accounts for "relative variance" or "coefficient of variation" rather than the observed variation. Therefore, inpatient variability relative to its mean is used as a measure, and in this case, no longer dominates the PCA loadings for PC I. Thus, PC I represents a weighted sum of expenditures, with the greatest weight assigned to office-based visits and medications, which may be important due to their relatively common usage compared to other types of expenditures. So, if PC I represents a measure of the "size" of medical expenditures, the other Principal Components represent the "shape" of health care use. We see in the PC II after log-transformation, positive loadings for vision and, especially, dental, and relatively small negative loadings for other health care uses. Therefore, we can assume that there are a secondary group of people who have high dental/vision expenditures and low levels of other medical costs, and vice versa. We believe that applying Principal Component Analysis to complex survey data can provide insights into multivariate data, and also a helpful way to visualize the population. Using calculated principal components in PROC SURVEYREG (MODEL PC2=PC1), is a handy way to plot principal components and understand the areas of population density. Display 6 illustrates that the highest density of individuals is at the intercept, indicating most people 20-65 have relatively low costs. The darker shaded regions with positive PC II values reflect those with relatively more dental/vision costs.

EXPLORATORY FACTOR ANALYSIS

Another multivariate technique with similar processes but different aims than Principal Component Analysis is Exploratory Factor Analysis (EFA), which utilizes PROC FACTOR in SAS. Often, EFA starts with PCA, then rotates the dimensions, generally to be more equivalent among the variability accounted for, rather than the variability accounted by PC I exceeding that of all other PCs. In addition, the rotation can make the resulting dimensions more "interpretable" than PCA results. Both EFA and PCA have their proponents, but in general, PCA is utilized for dimension reduction and understanding the nature of the relationships among variables, whereas EFA assumes there are underlying latent and

	Matrix Used for Principal Component Analysis					
	Correlation Matrix		Covariance Matrix		Covariance Matrix of Log-Transformed Variables	
Medical Expense Type	PRIN1 (17.9%)	PRIN2 (12.0%)	PRIN1 (49.2%)	PRIN2 (19.3%)	PRIN1 (37.1%)	PRIN2 (15.6%)
Office-Based Visits	0.533	-0.245	0.196	0.699	0.639	-0.004
Medication Costs	0.279	0.379	0.087	0.482	0.622	-0.215
Outpatient Stays	0.500	-0.259	0.174	0.453	0.260	-0.126
Emergency Visits	0.248	0.006	0.022	0.024	0.155	-0.282
Inpatient Stays	0.475	0.123	0.961	-0.269	0.141	-0.284
Medical Supplies	0.246	0.190	0.006	0.005	0.064	-0.037
Home Health	0.159	0.663	0.019	0.009	0.028	-0.040
Vision	0.067	-0.373	0.000	0.002	0.103	0.120
Dental	0.103	-0.316	0.002	0.012	0.280	0.872

Table 3. First two eigenvectors for three approaches to Principal Component Analysis of the MEPS 2016 types of expenditures



Display 6. Plot of the first two Principal Components from the analysis of the log-transformed covariance matrix of nine types of health care expenditures.

unmeasurable factors, and the observed measurements can identify these factors based on their correlation structure. A common application is scale development based on a questionnaire, such as measuring general or disease specific quality of life, functional status, and symptom clusters. EFA of the variables or "instances" within an instrument provides a framework for how many domains are the basis for the scales, and which observed variables are correlated with them.

We demonstrate the use of EFA in complex survey data with the twelve questions from the SF-12 in the MEPS 2016 data set. The questions' numeric directionality was transformed so higher values represent the best health or least limitations. The %SURVEYCORRCOV macro call for the creation of the SAS correlation data set variables for complex survey data, and the subsequent three factor analysis rotations (non-rotated, orthogonal, and oblique) is:

```
title1 "Using the %SURVEYCORRCOV matrix to produce the";
title2 "correlation matrix of twelve questions from";
title3 "MEPS 2016 data. Ages 20-65";
 %SurveyCorrCov(data=a,strata=VARSTR, cluster=VARPSU,
        weight=PERWT16F, domain=include, subGrp=1, outp=SF12corr,
        varlist=ADCAPE42 ADCLIM42 ADDAYA42 ADDOWN42 ADGENH42
              ADMALS42 ADMWLM42 ADPAIN42 ADPALS42 ADPWLM42 ADNRGY42
              ADSOCA42);

title1 "EFA based on the corr matrix; no rotation";
proc factor data=SF12corr(TYPE=CORR) method=principal score ;
  var ADCAPE42 ADCLIM42 ADDAYA42 ADDOWN42 ADGENH42 ADMALS42
        ADMWLM42 ADPAIN42 ADPALS42 ADPWLM42 ADNRGY42 ADSOCA42;

title1 "EFA based on the corr matrix; orthogonal rotation";
proc factor data=SF12corr(TYPE=CORR) rotate=varimax score ;
  var ADCAPE42 ADCLIM42 ADDAYA42 ADDOWN42 ADGENH42 ADMALS42
        ADMWLM42 ADPAIN42 ADPALS42 ADPWLM42 ADNRGY42 ADSOCA42;

title1 "EFA based on the corr matrix; non-orthogonal rotation";
proc factor data=SF12corr(TYPE=CORR) rotate=promax score;
  var ADCAPE42 ADCLIM42 ADDAYA42 ADDOWN42 ADGENH42 ADMALS42
        ADMWLM42 ADPAIN42 ADPALS42 ADPWLM42 ADNRGY42 ADSOCA42;
```

The results are summarized in Table 4. Two factors result based on the SAS default method, using eigenvalue greater than "1." This is consistent with the two component scores (Physical Component Score, PCS, and Mental Component Score, MCS) that are used in practice for this instrument. As noted in the previous section, once again we see the common pattern for the unrotated factors. All twelve variables are positively correlated, so the first factor loadings are all positive, indicating collectively a "size" measure of health status. In contrast, the second factor captures the "shape" of health status. The second factor indicates the second source of variability contrasts "physical limitations" (particularly the five most negative loadings) with "mental health limitations" (particularly the five most positive loadings).

The second PROC FACTOR rotates these factors, so they remain uncorrelated (orthogonal), with a VARIMAX rotation. Table 4 shows the transformation from unrotated to rotated: Factor I increased the positive loading for the "physical limitations" components, and transformed the "mental health" to relatively small negative loadings; Factor II is vice versa. Therefore, an individual with high scores for Factor I might represent "physical health," and Factor II high scores may indicate "mental health." But it must be noted that these are "relative" scores, because they contrast one another with non-trivial negative

factor loadings in two variables each. Therefore, the highest scores in Factor I are individuals with high scores in the physical activities questions, and low scores for being calm, peaceful and not depressed (i.e., no physical limitations but indications of poorer mental health). This represents the consequence of the overall positive correlations among the twelve questions: to maintain uncorrelated factors, some variables must maintain negative loadings. In general, the positive correlations for the overall population indicate that this combination of good physical measures and poor mental health measures is the exception to the rule. However, in specific disease states, the correlation structure may differ from the US general population. Therefore, whether using MCS or PCS, or the two orthogonal factors in Table 3, these cannot be interpreted singly but only in combination. For instance, a decrease in Factor I over time in a subgroup needs to be interpreted in combination with Factor II, because "Factor I: physical health" decreases when physical limitations remain the same and mental health improves.

Therefore, even though orthogonal rotations have benefits such as measuring unique constructs, the oblique (or correlated) Promax rotation results within Table 4 contain relatively trivial loadings for negative coefficients. Therefore, a decrease in mental health will not adversely impact the "physical health" score of Factor I. Thus, one must make the choice when variables are all positively correlated, a common occurrence: either maintain uncorrelated factors with negative loadings or accept that the two factors are correlated with oblique rotations.

	Type of Factor Analysis Rotation					
	Unrotated Principal Components		Orthogonal VARIMAX Rotation		Oblique PROMAX Rotation	
SF-12 Question	FACT1	FACT2	FACT1	FACT2	FACT1	FACT2
Lack of Limitations of Moderate Activities	0.118	-0.313	0.294	-0.161	0.237	-0.076
Lack of Limits Climbing Stairs	0.118	-0.289	0.278	-0.143	0.227	-0.063
Less Work Limitations	0.133	-0.239	0.257	-0.095	0.220	-0.022
Did Not Accomplish Less due to Physical Problems	0.129	-0.197	0.226	-0.065	0.199	-0.002
Pain Does Not Limit Work	0.118	-0.164	0.196	-0.048	0.175	0.007
General Health	0.104	0.001	0.079	0.068	0.095	0.087
Mental Problems Does Not Limit Work	0.117	0.099	0.025	0.151	0.066	0.152
Health Does Not Stop Social Activities	0.121	0.172	-0.020	0.209	0.039	0.196
Did Not Accomplish Less due to Mental Problems	0.121	0.193	-0.034	0.225	0.030	0.208
Lots of Energy	0.104	0.219	-0.064	0.234	0.004	0.208
Do Not Feel Depressed	0.101	0.362	-0.159	0.341	-0.057	0.285
Felt Calm Peaceful	0.085	0.386	-0.186	0.349	-0.082	0.285

Table 4. The standardized scoring coefficients for two factors resulting from Exploratory Factor Analysis for three rotations of the MEPS 2016 SF-12 questions.

CANONICAL CORRELATION

Canonical correlation analysis is often covered in multivariate courses, but perhaps underutilized in practice. Canonical correlation takes two sets of variables and finds a set of coefficients that maximize their correlation. That is, compared to multivariable regression, which maximizes the correlation of the linear combination of the independent variables with a single dependent variable, canonical correlation analysis maximizes correlation of a linear combination of one set of variables with a linear combination of a second set of variables. Canonical correlations can serve as a baseline to understand the relationships among sets of variables and may inform situations to determine whether one needs models for multiple outcome variables or a single omnibus model will suffice. For instance, rather than creating a regression model to estimate inpatient costs, and another regression model for outpatient costs, et cetera, canonical correlation may indicate what variables can be used for an overall model, and which dependent variables may be more unique, and in need of ~~a~~ their own model.

We demonstrate the use of canonical correlation in complex survey data again with the MPS 2016 data, with two sets of variables: the twelve questions from the SF-12 and the nine log-transformed types of medical costs. Oftentimes, there may be important covariates to adjust for, so we will include age in the PARTIAL statement in the second example. The first %SURVEYCORRCOV macro call generates the correlation matrix of the twelve questions and the nine log-transformed cost variables, and then PROC CANCORR performs the analysis to maximize the correlation between the SF-12 questions and types of medical expenses. If one is not interested in the multivariate tests such as Wilks' Lambda and only the canonical structure and correlations, this will suffice. However, if a test of Wilks' Lambda is of interest, adjusting the sample size with the option of the correlation matrix designed for PROC REG (OUTP4REG) will be beneficial. We randomly select one of the variables to be the "dependent" variable, and the design effect adjusts the N accordingly for Wilks' Lambda. We demonstrate this use of %SURVEYCORRCOV and the PROC CANCORR PARTIAL statement in the second example:

```
title1 "Using the %SURVEYCORRCOV matrix to produce the";
title2 "correlation matrix of nine types of expenses";
title3 "and twelve SF-12 questions from MEPS 2016 data";
%SurveyCorrCov(data=a, strata=VARSTR, cluster=VARPSU,
        weight=PERWT16F, domain=include, subGrp=1, outp=CanCorr1,
        varlist=ADCAPE42 ADCLIM42 ADDAYA42 ADDOWN42 ADGENH42
             ADMALS42 ADMWLM42 ADPAIN42 ADPALS42 ADPWLM42
             ADNRGY42 ADSOCA42 logDental logER logHome logOffice
             logOutPt logRX logInPt logOthSup logVision);

title1 "Canonical Correlation, No Sample Size Adjustment";
proc cancorr data=CanCorr1(TYPE=CORR);
        var ADCAPE42 ADCLIM42 ADDAYA42 ADDOWN42 ADGENH42
             ADMALS42 ADMWLM42 ADPAIN42 ADPALS42 ADPWLM42
             ADNRGY42 ADSOCA42;
        with logDental logER logHome logOffice
             logOutPt logRX logInPt logOthSup logVision;run;

title1 "Using the %SURVEYCORRCOV matrix to produce the";
title2 "correlation matrix of nine types of expenses";
title3 "twelve SF-12 questions, age from MEPS 2016 data";
%SurveyCorrCov(data=a, strata=VARSTR, cluster=VARPSU,
        weight=PERWT16F, domain=include, subGrp=1,
             outp4reg=CanCorr2,
```

```
        indep=ADCAPE42 ADCLIM42 ADDAYA42 ADDOWN42
              ADMALS42 ADMWLM42 ADPAIN42 ADPALS42 ADPWLM42
              ADNRGY42 ADSOCA42 logDental logER logHome logOffice
              logOutPt logRX logInPt logOthSup logVision age16X,
              depend=ADGENH42);

title1 "Partial Canonical Correlation, Sample Size Adjustment";
proc cancorr data=CanCorr2(TYPE=CORR);
        var ADCAPE42 ADCLIM42 ADDAYA42 ADDOWN42 ADGENH42
              ADMALS42 ADMWLM42 ADPAIN42 ADPALS42 ADPWLM42
              ADNRGY42 ADSOCA42;
        with logDental logER logHome logOffice
              logOutPt logRX logInPt logOthSup logVision;
        partial age16X; run;
```

The canonical correlation coefficients both with and without age adjustment are presented in Table 5. The five SF-12 questions about pain limiting work, general health, limits climbing stairs, work limitations, and energy were most related to medical expenses. Individuals with higher (better health) for those SF-12 items tended to have lower medical costs, particularly medication costs, as indicated by their negative loadings: as self-reported health was better, expenses were lower. The exceptions are dental care, which had a positive loading, and vision care with a smaller positive loading. Therefore, those with better health tended to have higher dental expenses, which may be related to socioeconomic factors related to access to dental insurance and advanced dental procedures.

	Standardized Canonical Coefficient			Standardized Canonical Coefficient	
SF-12 Question	No Age Adj	Age Adj	Type of Medical Expense	No Age Adj	Age Adj
Pain Does Not Limit Work	0.289	0.276	Medication Costs	-0.645	-0.609
General Health	0.267	0.264	Home Health	-0.244	-0.264
Lack of Limits Climbing Stairs	0.226	0.199	Medical Supplies	-0.207	-0.214
Less Work Limitations	0.202	0.187	Emergency Visits	-0.164	-0.208
Lots of Energy	0.198	0.213	Outpatient Stays	-0.133	-0.120
Lack of Limitations of Moderate Activities	0.084	0.097	Office-Based Visits	-0.132	-0.119
Did Not Accomplish Less Due to Mental Problems	0.035	0.072	Inpatient Stays	-0.124	-0.148
Health Does Not Stop Social Activities	0.028	0.075	Vision	0.080	0.083
Do Not Felt Depressed	0.024	0.017	Dental	0.239	0.279
Did Not Accomplish Less Due to Physical Problems	0.012	-0.003			
Mental Problems Does Not Limit Work	-0.025	-0.022			
Felt Calm Peaceful	-0.075	-0.075			

Table 5. The canonical coefficients maximizing the correlation between the log-transformed health care cost categories and the questions of the SF-12, using the MEPS 2016 data. Results are shown with and without adjustment for age.

CONCLUSION

Complex survey data is an important segment of real-world evidence, often providing a bounty of high quality, unique information at no cost. However, this data must be analyzed properly to account for the survey design, and not as a simple random sample. %SURVEYCORRCOV both replicates survey-based analyses in PROC CORR and is a conduit to other procedures. We hope this macro helps broaden the types of analysis performed with complex survey data. We encourage any comments/questions about problems, suggestions, and experiences with the macro, particularly for uses in other SAS PROCs such as structural equation modeling (PROC CALIS). Updates and current versions of the %SURVEYCORRCOV macro are maintained at:

https://github.com/DavidRNelson/-surveycorrcov-sas-macro.

REFERENCES

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J. and Münkemüller, T., 2013. "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance." *Ecography*, 36:27-46.

Harrell Jr, F.E., 2015. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. New York, NY: Springer.

Heeringa, S.G., West, B.T. and Berglund, P.A., 2017. *Applied survey data analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Hampton, J. 2013. "Proc SurveyCorr." Accessed February 1, 2020. https://www.lexjansen.com/nesug/nesug13/34_Final_Paper.pdf.

Lewis, T.H., 2016. *Complex survey data analysis with SAS*. Boca Raton, FL: Chapman and Hall/CRC.

Nelson, B.D., 2001. "Variable Reduction for Modeling using PROC VARCLUS." Accessed February 11, 2020. https://support.sas.com/resources/papers/proceedings/proceedings/sugi26/p261-26.pdf.

Wang, F. and Shin, H.C., 2011. "SAS® Model Selection Macros for Complex Survey Data Using PROC SURVEYLOGISTIC/SURVEYREG." Accessed February 10, 2020. http://www.mwsug.org/proceedings/2011/stats/MWSUG-2011-SA02.pdf.

Ware Jr, J.E., Kosinski, M. and Keller, S.D., 1996. "A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity." *Medical Care*, 34:220-233.

West, B.T., Sakshaug, J.W. and Aurelien, G.A.S., 2016. "How big of a problem is analytic error in secondary analyses of survey data?" *PloS one*, 11(6).

ACKNOWLEDGMENTS

We appreciate the input and suggestions of Anthony Zagar and Jiat-Ling Poon.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

David R. Nelson
Eli Lilly & Company
nelson_david_r@lilly.com