Paper PO-5

SAS® Macros to Conduct Common Biostatistical Analyses and Generate Reports

Dana Nickleach, Yuan Liu, Adam Shrewsberry, Kenneth Ogan, Sungjin Kim, and Zhibo Wang, Emory University, Atlanta, GA

ABSTRACT

Minimize your time spent on common biostatistical analyses by maximizing your use of these macros. Put them to work calculating statistics and producing high quality report tables summarizing your results in Word documents. These macros are useful for conducting a complete analysis, from start to finish. Use them to 1) produce descriptive statistics, including frequencies and percentages for categorical variables; and n, mean, median, standard deviation, min, and max for quantitative variables; 2) produce parametric and non-parametric bivariate statistics with either quantitative or categorical variables, including Chi-Square test, Fisher's exact test, ANOVA, Kruskal-Wallis test, Pearson correlation coefficient, and Spearman rank correlation coefficient dependent on variable types; 3) look at the unadjusted associations of each variable with a binary or survival outcome, reporting odds ratios or hazard ratios, respectively; 4) conduct multiple regression using logistic regression or Cox proportional hazards models incorporating all variables or using a backward variable selection method. The capabilities of these macros and how to use them will be illustrated using data from a kidney stone questionnaire designed to examine the factors that influence patient preference for ureteroscopy vs. shock wave lithotripsy. The macros used for this study enabled production of comprehensive, professional looking reports, efficient communication and collaboration with investigators, and ensured that timely and high-quality service was delivered.

INTRODUCTION

Our group of biostatisticians set out to produce high quality, professional looking analysis reports; enhance communication and collaboration with investigators or clients; and also save time on common tasks. We ended up developing a set of macros to accomplish those goals. The majority of these macros can be useful for general statistical analysis and are not specific to biostatistics. They can take you through the entire analysis process from start to finish, including descriptive statistics, bivariate statistics or unadjusted regression, and multiple regression. The macros allow the user to comprehensively examine variables and associations with ease. Raw SAS output can be cumbersome and may not be understandable to the non-statistician or non-SAS user. These macros generate tables with the goal of condensing the relevant information from a statistical analysis into an understandable format for investigators. Each macro produces a table in a rich text format (RTF) file containing the results.

We illustrate the use of these macros using data from a cross-sectional study of patient preference in treatment of urolithiasis. A self-completed questionnaire was collected from patients with urolithiasis in our stone clinic. The main goal of the study was to examine patient preference for shock wave lithotripsy (SWL) vs. ureteroscopy (URS) and the factors that influence their preference. The outcome was the binary patient preference variable. The main predictors of interest were the decision making factors, which included success, complications, need for stent, and need for 2nd surgery. Demographics and kidney stone history were examined as covariates. Demographics included age, gender, body mass index (BMI), race, education, marital status, household income, and employment. Information on kidney stone history included family history, first stone, number of kidney stones, age of first stone, time elapsed since last stone, number of kidney stone surgeries, surgery complications, number of ER visits due to kidney stones, number of work days missed due to stones, dietary changes, and stone medications.

Before proceeding with the macros there are several things to point out. Many of the macros refer to categorical and numerical variables. Categorical variables should have a discrete number of categories and can be a character or numeric variable. Numerical variables will be treated as quantitative variables and must be a numeric variable type. The macros cover several outcome or dependent variable types including, binary, time to event, and quantitative. Some macros will not work correctly with variable names that are longer than 20 characters. We recommend renaming variables so that they meet this requirement before using the macros. All macros have a DEBUG parameter. Set it to T to run the macro in debug mode. In debug mode, temporary data sets created by the macro will not be deleted. By default, the data sets will be deleted at the end of the macro in order to keep the work library clean. Debug mode is useful if you are trying to debug the code, editing the code, or want to further manipulate the resulting data sets. Checking for errors in some macro parameters has also been implemented as described previously (Karafan, 2011).

DESCRIPTIVE STATISTICS

The first step in almost any analysis is to look at descriptive statistics. Our first macro, %DESCRIPTIVE, produces a table including frequencies and percentages for categorical variables; and n, mean, median, standard deviation, min, and max for quantitative variables. This has several purposes, first, to scan the data. Make sure it looks correct and

all values make sense. For example, there should not be a maximum age of 120. The number of missing values is also reported, which may indicate some variables may not be usable due to a high number of missing values or methods may need to be employed to handle them. Also, some categories that have small frequencies may need to be collapsed when looked at in further analysis. Secondly, it is important to know who was sampled. A table that describes the sample is a very standard table in reports and manuscripts. The parameters for the %DESCRIPTIVE macro are presented in Table 1.

Parameter	Description	Required
DATASET	The name of the data set to be analyzed.	Yes
CLIST	List of categorical variables, separated by empty space.	Yes
NLIST	List of numerical variables, separated by empty space.	Yes
OUTPATH	File path for output table to be stored.	Yes
FNAME	File name for output table.	Yes
FOOTNOTE	Text to include in the footnote of the table. It should be in quotes. Leave this field blank if not including a footnote.	No
CHI	Set to T to calculate the chi-square goodness of fit p-value for categorical variables. This is useful if you are interested in whether or not categories have equal proportions. The default value is F.	No

Table 1. Parameters for %DESCRIPTIVE Macro

We used this macro to describe all 30+ variables in the kidney stone data set. However, in the interest of space, the following example only uses 4 of those variables. Figure 1 is a screenshot of the RTF file that is produced. Categorical and numerical variables are summarized together in one table. The variables appear in the table in the same order that they are listed in the macro call. However, all numerical variables will always appear after all categorical variables. Variable labels are always displayed if available. Be aware that the macro will not run correctly if more than one variable has the same label unless one is specified in CLIST and the other in NLIST.

```
TITLE 'Table 1 Descriptive Statistics';
%DESCRIPTIVE(DATASET=anal,
       CLIST = sex race_,
       NLIST = AgeYears_ bmi,
       OUTPATH = C:\Users\Dana\Documents\,
       FNAME = Table 1 Descriptive Statistics,
       FOOTNOTE ='Abbreviations: BMI - Body Mass Index.');
TITLE;
```

Table 1 Descriptive Statistics

Variable	Level	N = 163	%
Sex	Female	67	41.1
	Male	96	58.9
Race	White	130	80.2
	Other	32	19.8
	Missing	1	-
Age (yrs)	Mean	53.72	-
	Median	55	-
	Minimum	18	-
	Maximum	82	-
	Std Dev	14.97	-
	Missing	2	-
BMI	Mean	28.90	-
	Median	27.06	-
	Minimum	16.64	-
	Maximum	61.79	-
	Std Dev	6.82	-
	Missing	4	-

Abbreviations: BMI - Body Mass Index.

Figure 1. Output Produced by %DESCRIPTIVE Macro

BIVARIATE STATISTICS

After getting an initial feel for the data, we begin to explore associations of the covariates with variables of interest, i.e. outcomes and main predictors. The macro %UNI_CAT can produce parametric and non-parametric bivariate statistics for a list of covariates with a categorical variable of interest. For categorical covariates, frequencies and percentages from a contingency table are reported; and the Pearson chi-square test and Fisher's exact test are conducted. For numerical covariates, means and medians are reported; and ANOVA and the Kruskal-Wallis test are conducted. The parameters for %UNI_CAT are presented in Table 2.

Parameter	Description	Required
DATASET	The name of the data set to be analyzed.	Yes
OUTCOME	Categorical variable to be associated with CLIST and NLIST variables. More than one variable can be listed separated by empty space. However, these variables appear in the table header and too many variables will cause the table to wrap due to the document page width limitations, producing undesirable results. Each variable name must not be more than 30 characters long.	Yes
CLIST	List of categorical variables, separated by empty space.	Yes
NLIST	List of numerical variables, separated by empty space.	Yes
OUTPATH	File path for output table to be stored.	Yes
FNAME	File name for output table.	Yes
NONPAR	Specify a value of F, T, or A to indicate whether to conduct non-parametric tests. If the value is T then both parametric and non-parametric tests will be conducted. If the value is F then only parametric tests will be conducted. A value of A means that for categorical variables, the appropriate test statistic, non-parametric or parametric, will be automatically chosen based on whether the chi-square test is invalid, but for numerical covariates only the parametric test will be calculated. Option A is only available for SAS V9.3 or later. The default value is F.	No
SPREAD	Set to T to also report standard deviation, min, and max for numerical variables. The default value is F.	No
BY	A separate analysis will be conducted for each value of the variable specified here.	No
MHC	Set to T to report p-values from Mantel-Haenszel chi-square tests instead of Pearson chi-square tests. The default value is F.	No
ROWPERCENT	Set to F to report column percentages instead of row percentages from the contingency table. The default value is T.	No
ORIENTATION	Value of PORTRAIT or LANDSCAPE to indicate the page layout of the report. The default value is PORTRAIT.	No
WEIGHT	Weight variable to use in a WEIGHT statement. Weights will not be normalized by the macro. The reported N will be the sum of the weights. This option will not work with NONPAR = T or A. Leave blank if not weighting.	No

Table 2. Parameters for %UNI_CAT Macro

The following code was used to explore the unadjusted association of each covariate with the outcome, procedure preference. Again, we are only using a handful of variables in the interest of space. Figure 2 displays the results that are produced. Significant p-values (p-value <.05) are automatically highlighted in bold font. After deciding that parametric p-values would be sufficient for the continuous variables, we set NONPAR=A so that Fisher's exact p-values would automatically be reported for categorical variables when the chi-square assumption was invalid. However, if you have some continuous variables that do not meet the assumptions for ANOVA, set NONPAR=T to get a column for both parametric and non-parametric p-values. We set ROWPERCENT=T since procedure preference is considered to be an outcome. If the variable is not considered to be an outcome then column percentages may make more sense. We also used this macro to look at the association of the other covariates with the main predictors, the decision making influences, which were also binary variables. However, if you are interested in associations with a numerical variable, use the macro %UNI_NUM. %UNI_NUM operates very similarly to %UNI_CAT. Pearson correlation and Spearman rank correlation coefficients along with p-values are reported for numerical covariates; and mean, median, ANOVA, and Kruskal-Wallis p-values are reported for categorical covariates.

```
TITLE 'Table 4 Unadjusted Associations of Demographics and Stone History with
Preference';
%UNI_CAT(DATASET=anal,
       OUTCOME=ProcedurePreference_,
       CLIST=sex race_,
       NLIST=AgeYears_ bmi,
       NONPAR=A,
       SPREAD=T,
```

```
        OUTPATH=C:\Users\Dana\Documents\,
        FNAME=Table 4 Unadjusted Associations of Demographics and Stone History with
        Preference,
        ROWPERCENT=T);
TITLE;
```

Table 4 Unadjusted Associations of Demographics and Stone History with Preference

			Procedure Preference		
Covariate	**Statistics**	**Level**	**URS N=102**	**SWL N=61**	**P-value***
Sex	N (Row %)	Female	47 (70.15)	20 (29.85)	0.095
	N (Row %)	Male	55 (57.29)	41 (42.71)	
Race	N (Row %)	White	76 (58.46)	54 (41.54)	**0.017**
	N (Row %)	Other	26 (81.25)	6 (18.75)	
Age (yrs)	N		100	61	0.762
	Mean		53.44	54.18	
	Median		55	56	
	Min		18	20	
	Max		82	81	
	Std Dev		15.05	14.97	
BMI	N		99	60	0.065
	Mean		29.68	27.62	
	Median		27.37	26.54	
	Min		18.37	16.64	
	Max		61.79	41.84	
	Std Dev		7.62	5.06	

* The p-value is calculated by ANOVA for numerical covariates; and chi-square test or Fisher's exact for categorical covariates, where appropriate.

Figure 2. Output Produced by %UNI_CAT Macro

UNADJUSTED REGRESSION

We can continue to explore unadjusted associations of covariates with the outcome using a regression model. We conduct unadjusted logistic regression using the macro %UNI_LOGREG, reporting odds ratios. Note that for a binary outcome, it is not necessary to run both %UNI_CAT and %UNI_LOGREG. Most of the time investigators prefer to see percentages as produced by %UNI_CAT instead of odds ratios as produced by %UNI_LOGREG. However, in some cases %UNI_NUM and %UNI_CAT are not appropriate options. For example, if you have a time to event outcome you will need to conduct Cox proportional hazards regression to examine the unadjusted associations. The parameters for %UNI_LOGREG are presented in Table 3.

Parameter	Description	Required
DATASET	The name of the data set to be analyzed.	Yes
OUTCOME	The name of the binary or ordinal response variable.	Yes
EVENT	The event category for the binary response model. You can specify the value in quotes. This will be passed to the EVENT= option in the MODEL statement. Leave this blank if the response is ordinal and not binary.	No
CLIST	List of categorical variables, separated by empty space.	Yes
CREFLIST	List of the categorical variables with reference levels specified as in a CLASS statement, but with variables separated by an asterisk. For example,	No

	CREFLIST= var1 (REF='A') * var2 (DESC). The order of the variables needs to be the same as in CLIST.	
NLIST	List of numerical variables, separated by empty space.	Yes
TYPE3	Set to F to suppress type III p-values from being reported in the table. The default value is T.	No
OUTPATH	File path for output table to be stored.	Yes
FNAME	File name for output table.	Yes
ORIENTATION	Value of PORTRAIT or LANDSCAPE to indicate the page layout of the report. The default value is PORTRAIT.	No

Table 3. Parameters for %UNI_LOGREG Macro

The code below fits a separate logistic regression model for each variable listed in CLIST and NLIST with the OUTCOME variable used as the response variable. The output produced appears in Figure 3. Sample sizes, odds ratios, 95% confidence intervals, odds ratio p-values, and type III p-values are reported. The response that is modeled is included in the table header. If necessary, reference groups can also be specified. Type III p-values will be equivalent to the odds ratio p-values if all categorical variables have only 2 levels or only numeric variables are used. Therefore, type III p-values can be suppressed using the TYPE3 parameter to avoid confusion.

```
TITLE 'Table 4 Unadjusted Logistic Regressions of Demographics and Stone History with
Preference';
%UNI_LOGREG(DATASET=anal,
       OUTCOME=ProcedurePreference_,
       EVENT='SWL',
       CLIST=sex SWL_URS,
       NLIST=AgeYears_ bmi,
       OUTPATH=C:\Users\Dana\Documents\,
       FNAME=Table 4 Unadjusted Logistic Regressions of Demographics and Stone History
with Preference);
TITLE;
```

Table 4 Unadjusted Logistic Regressions of Demographics and Stone History with Preference

Covariate	Level	N	Procedure Preference=SWL Odds Ratio	95%CI Low	95%CI Up	OR P-value	Type3 P-value
Sex	Female	67	0.571	0.295	1.106	0.097	0.097
	Male	96	-	-	-	-	
Previous URS/SWL	Both URS and SWL	44	0.833	0.354	1.963	0.677	**0.018**
	URS, but not SWL	35	0.403	0.146	1.112	0.079	
	SWL, but not URS	37	2.115	0.880	5.082	0.094	
	Neither URS nor SWL	47	-	-	-	-	
Age (yrs)		161	1.003	0.982	1.025	0.760	0.760
BMI		159	0.952	0.902	1.004	0.070	0.070

Figure 3. Output Produced by %UNI_LOGREG Macro

%UNI_PHREG is a similar macro that works with a time to event outcome and fits Cox proportional hazards models. Hazard ratios instead of odds ratios are reported. Additionally, the proportional hazards assumption can be tested. Another macro, %UNI_GENMOD uses PROC GENMOD and will work with a quantitative outcome. If a normal distribution is specified then the coefficient estimates are reported. If a Poisson or negative binomial distribution is specified, a log link function is used, and the rate ratio is reported. An option is also available to specify the use of generalized estimating equations (GEE) in that macro if you have repeated measures.

MULTIPLE REGRESSION

Finally, we are ready to conduct multiple regression. The results from the previous macros helped to prepare us for this step and give us an idea of which variables to consider as candidates in the model. Depending on the situation we may decide to consider only confounders or all covariates as candidates for the model. However, the user should be cognizant of possible colinearity issues when selecting variables. We will use a variable selection method to select the final model covariates. The macro %LOGREG_SEL will conduct backward selection on a logistic regression model. The maximum possible sample size at each stage of the selection process will be used instead of restricting to the sample size from the first step as SAS does when using their selection methods. The final model results will be reported in a table if requested. The parameters for %LOGREG_SEL are presented in Table 4. The macro %MULTIPLE_LOGREG must be compiled before using %LOGREG_SEL.

Parameters	Description	Required
DSN	The name of the data set to be analyzed.	Yes
OUTCOME	The name of the binary or ordinal response variable.	Yes
EVENT	The event category for the binary response model. Specify the value in quotes. This is the argument that will be passed to the EVENT= option in the MODEL statement. Leave this blank if the response is ordinal and not binary.	No
DESC	Set to T to reverse the order of the response variable. The ordering will be based on the internal sort order. Only specify this if the EVENT parameter is blank. The default value is F.	No
VAR	List of all variables to include in the model separated by spaces. Each variable name or interaction term must not be more than 20 characters long.	Yes
CVAR	List of categorical variables to include in the model separated by spaces. These should also appear in the VAR parameter. To change the reference group by reversing the sort order of the categories, follow each variable name by the (DESC) option where desired. However, separate variables with an asterisk instead of a space.	Yes
INC	Number of variables to force in the model. The first *n* variables in the VAR parameter will be included in every model. The default value is 0.	No
SLSTAY	The significance level for removing variables from the model. The default value is .05.	No
WEIGHT	Variable to use in the weight statement. Weights will be normalized to the original sample size using the normalize option. Leave it blank if not using weights.	No
REPORT	Set this to T if you want a table of the final model generated. The default value is F.	No
TYPE3	Set to F to suppress type III p-values from being reported in the table. The default value is T.	No
OUTPATH	File path for output table to be stored.	Yes, if REPORT=T
FILENAME	File name for output table.	Yes, if REPORT=T

Table 4. Parameters for %LOGREG_SEL Macro

The following code was used to conduct the backwards selection. The first four variables, the decision making factors, are forced in the model since they are the main predictors of interest and therefore, should be reported regardless of significance. Setting INC=4 accomplishes this. Number of stone surgeries, previous complications, and number of stones in lifetime are not included in order to avoid mutlicollinearity issues and are not listed in the macro call. The rest of the covariates were entered into the model subject to removal from the model using an alpha=.20 removal criteria (SLSTAY=.2). REPORT=T was specified so that the results in Figure 4 are produced. The output is similar to %UNI_LOGREG. Additionally, the number of observations read and used by PROC LOGISTIC are reported in the footnote, as well as the list of variables removed from the model and the selection criteria. The final list of variables selected will also be saved in global macro variables and written to the log so that they can easily be copied and pasted. For greater flexibility, the user may choose to fit a model outside of the macro using the final variable list.

```
TITLE 'Table 6 Multiple Logistic Regression';
```

```
%LOGREG_SEL(DSN=anal,
       OUTCOME=ProcedurePreference_,
       EVENT='URS',
       VAR=Success2 Complications2 NeedForStent2 NeedForSurgery2 sex race_ income_
       education_ employment_ married_ FH_of_stones_ First_stone time_last SWL_URS
       PCNL_c Open_c Stent_c VisitsED Missed_work_days diet_typ Meds_for_stones_
       AgeYears_ bmi Age_of_first_stone_,
       CVAR=Success2 Complications2 NeedForStent2 NeedForSurgery2 sex race_ income_
       education_ employment_ married_ FH_of_stones_ First_stone time_last SWL_URS
       PCNL_c Open_c Stent_c VisitsED Missed_work_days diet_typ Meds_for_stones_,
       INC=4,
       SLSTAY=.2,
       REPORT=T,
       OUTPATH=C:\Users\Dana\Documents\,
       FILENAME=Table 6 Multiple Logistic Regression);
TITLE;
```

Table 6 Multiple Logistic Regression

Covariate	Level	Procedure Preference=URS				
		Odds Ratio	95%CI Low	95%CI Up	OR P-value	Type3 P-value
Success	Large/Extremely	16.80	1.46	193.20	**0.024**	**0.024**
	Not at all/small/moderate	-	-	-	-	
Complications	Large/Extremely	0.24	0.07	0.84	**0.026**	**0.026**
	Not at all/small/moderate	-	-	-	-	
Need for Stent	Large/Extremely	0.38	0.17	0.85	**0.018**	**0.018**
	Not at all/small/moderate	-	-	-	-	
Need for 2nd Surgery	Large/Extremely	3.81	1.53	9.49	**0.004**	**0.004**
	Not at all/small/moderate	-	-	-	-	
Previous URS/SWL	Both URS and SWL	0.58	0.18	1.94	0.377	**0.022**
	URS, but not SWL	1.69	0.46	6.19	0.426	
	SWL, but not URS	0.25	0.08	0.83	**0.023**	
	Neither URS nor SWL	-	-	-	-	
Previous Stent Placement	Yes	2.25	0.89	5.67	0.085	0.085
	No	-	-	-	-	
BMI		1.07	0.99	1.15	0.074	0.074

* Number of observations in the original data set = 163.
Number of observations used = 154.
Backward selection with an alpha level of removal of .2 was used. The following variables were removed from the model: Age (yrs), Age of First Stone, Education, Employment, Family History of Stones, First Time with Stones, Income, On Medications for Stone Prevention, Number of Missed Work Days in Last Year Secondary to Stones, Previous Open Surgery, Previous PCNL, Sex, Number of ER Visits for Stones, Type of Dietary Changes for Stone Prevention, Married, Race, and Time Since Last Stone (Yrs).

Figure 4. Output Produced by %LOGREG_SEL Macro

If backward selection is not necessary, the user can alternatively fit the model themselves with PROC LOGISTIC in combination with ODS OUTPUT and then use the macro %MULTIPLE_LOGREG to produce a similar output table.

For time to event outcomes, there are similar macros %PHREG_SEL and %MULTIPLE_PHREG that fit Cox proportional hazards models. Additionally, a STRATA variable can be specified for these models. For a quantitative outcome, %MULTIPLE_LINREG will produce a table displaying the results from a general linear model after the user runs a model using PROC GLM. %MULTIPLE_GENMOD can be used for generalized linear models. A log link function is assumed with a binomial, Poisson, or negative binomial distribution; and an identity link with a normal distribution. Relative risk is reported for a binomial distribution, rate ratio for Poisson or negative binomial, and coefficients for normal models. The modeling must be done using PROC GENMOD before using this macro. This macro is also set up to handle the scenario of a GEE model when a REPEATED statement is used. When an interaction term is in a model, the backward selection macros will remove variables with respect to model hierarchy. However, none of these macros are able to correctly produce a table in this scenario.

CONCLUSION

These macros can cover every step of a common analysis, handling a variety of data types, and producing streamlined report tables. We do not claim that the tables produced are publication quality and do not expect all the information to be reported. However, the tables provide information with the purpose of understanding the data and conveying that information to an investigator. The investigator can then choose the story they want to tell and will have all the information available to extract what is needed. We have found that the macros have enabled efficient communication and collaboration with investigators, while also saving us time.

The development of these macros has been a continuous process. We have added to and revised the collection many times to make improvements, accommodate new needs, and increase flexibility. All macros and complete documentation are available on our website: https://winshipbbisr.emory.edu/Software_macro.html. Any feedback or bug reporting is greatly appreciated.

REFERENCES

Karafa, Matthew T. 2011. “Building Better Macros: Basic Parameter Checking for Avoiding "ID10T" Errors.” SAS Global Forum 2011 Conference Proceedings. Cary, NC: SAS Institute Inc. Available at: http://*support.sas.com/resources/papers/proceedings11/096-2011.pdf*.

CONTACT INFORMATION

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

Dana Nickleach
Biostatistics and Bioinformatics Shared Resource, Winship Cancer Institute, Emory University
1518 Clifton Rd., Room 5000O
Atlanta, GA 30322
Work Phone: 404-778-4874
E-mail:dnickle@emory.edu
Web: http://winshipbbisr.emory.edu