

1 ***%svy_logistic_regression*: A generic SAS® macro for simple and multiple logistic regression and**
2 **creating quality publication-ready tables using survey or non-survey data**

3 **Authors:** Jacques Muthusi ^{1*}, Samuel Mwalili ¹, Peter Young ¹

4 **Author affiliations:**

5 ¹ Division of Global HIV and Tuberculosis, U.S. Centers for Disease Control and Prevention, Nairobi,
6 Kenya

7 *** Corresponding author:**

8 Email address: mwj6@cdc.gov (JM)

9 **Author Contribution**

10 JM and SM took part in concept development. JM developed and documented the SAS macro, and
11 prepared the final manuscript. SM tested and debugged the SAS macro. PY helped define user
12 requirements and tested the SAS macro. All authors read and approved of the final manuscript for
13 publication.

14 **Keywords:** SAS macro, odds ratio, logistic regression, nutrition survey, data reporting, reproducible
15 research

16 **Competing interests:** The authors have declared that no competing interests exist.

17

18 **Abstract**

19 **Introduction:** Reproducible research is increasingly gaining interest in the research community.
20 Automating the production of research manuscript tables from statistical software can help increase the
21 reproducibility of findings. Logistic regression is used in studying disease prevalence and associated
22 factors in epidemiological studies and can be easily performed using widely available software including
23 SAS, SUDAAN, Stata or R. However, output from these software must be processed further to make it
24 readily presentable. There exists a number of procedures developed to organize regression output, though
25 many of them suffer limitations of flexibility, complexity, lack of validation checks for input parameters,
26 as well as inability to incorporate survey design.

27 **Methods:** We developed a SAS macro, *%svy_logistic_regression*, for fitting simple and multiple
28 logistic regression models. The macro also creates quality publication-ready tables using survey or non-
29 survey data which aims to increase transparency of data analyses. It further significantly reduces turn-
30 around time for conducting analysis and preparing output tables while also addressing the limitations of
31 existing procedures.

32 **Results:** We demonstrate the use of the macro in the analysis of the 2013-2014 National Health and
33 Nutrition Examination Survey (NHANES), a complex survey designed to assess the health and nutritional
34 status of adults and children in the United States. The output presented here is directly from the macro and
35 is consistent with how regression results are often presented in the epidemiological and biomedical
36 literature, with unadjusted and adjusted model results presented side by side.

37 **Conclusions:** The SAS code presented in this macro is comprehensive, easy to follow, manipulate and
38 to extend to other areas of interest. It can also be incorporated quickly by the statistician for immediate
39 use. It is an especially valuable tool for generating quality, easy to review tables which can be incorporated
40 directly in a publication.

41 **Introduction**

42 The principles of reproducible research are increasingly gaining interest both in the research community
43 (1-5) and in the popular imagination as a result of high-profile failures to reproduce results. While funders
44 and journals are increasingly requiring both publications and their supporting data be made publicly
45 available with few exceptions (6-8) there has been less focus on the reproducibility of the analysis process
46 itself. Reproducible research refers to increasing the transparency of the research endeavor by making the
47 initial data, detailed analysis steps, and tools available to allow others to reproduce ones' findings. Peng
48 and Leek refer to increasing reproducibility as a tool to reduce the time required to uncover errors in
49 analysis (9). One important link in the reproducible research value chain is eliminating manual
50 reformatting of results from statistical software into draft manuscript tables.

51 In most epidemiological studies, one of the main outcomes of interest is disease prevalence – i.e. the
52 proportion of all study subjects with a disease. Researchers are often interested in the probability or odds
53 of subjects having a disease as well as associated predictive factors. These factors can be categorical (such
54 as gender), ordinal (for example age categories), or continuous (for instance duration on treatment). The
55 measures of association are often presented as crude (unadjusted) odds ratios from simple logistic
56 regression or they can be presented as adjusted odds ratios from multiple logistic regression. In scientific
57 reports from observational epidemiological studies it is common to combine the results from multiple
58 statistical models and present the odds ratios side by side in complex tables showing the association
59 between multiple covariates with the outcome of interest, both unadjusted and adjusted.

60 Logistic regression models can be fitted easily using available standard statistical analysis software such
61 as SAS, SUDAAN, Stata or R, among others, and have been extended to handle weights and/or specialized

62 variance estimation to account for complex survey designs. However, output from these software is not
63 formatted for use directly in a publication and must be re-organized in order to make it more presentable
64 based on the cultural norms of the biomedical literature or the specific requirements of the scientific
65 journal (10, 11). Most epidemiological publications present regression tables showing odds ratios
66 estimates and the corresponding 95% confidence intervals and/or p-values. They further enrich the output
67 by including frequencies and proportion of study subjects who experienced the outcome of interest.
68 Results from simple and multiple regression can also be presented side-by-side in one table. Some
69 examples of publications which adopt this convention of presenting regression results are provided in the
70 references (12-15). In order to accomplish this, one has to manually copy different parts of output into a
71 template. This is both time-consuming and potentially prone to errors when revisions to the analysis are
72 required.

73 A number of programs have been developed to facilitate conversion of regression output from statistical
74 analysis software into formatted tables for publications. In Stata, several programs including *esttab* (16,
75 17), *reformat* (18), *outreg2* (19) are useful in formatting regression output. In R such packages as
76 *stargazer* (20), *broom* (21), *flextable* (22) have also being found helpful. Though they are useful to
77 statisticians they suffer from numerous limitations. For instance, they cannot automatically combine
78 results from several simple logistic regression into a single table. It is also not possible to combine results
79 from simple and multiple logistic regression into one output table. They are also not fully generic in that
80 one has to explicitly specify variable labels and levels of categorical variables instead of extracting these
81 from metadata. Further manipulation of output, for example, concatenating odds ratios and the
82 corresponding 95% confidence interval into one column cell, has to be done manually which increases the
83 risk of typographical errors in the output table. In SAS software, logistic regression models can be fitted
84 using the LOGISTIC, GENMOD and SURVEYLOGISTIC procedures (23), though output from these

85 procedures must be formatted further to make it presentable. SAS provides a flexible and powerful macro
86 language that can be utilized to create and populate numerous table templates for presenting regression
87 results. However, limited programming work has being done in SAS to date. There are several macros
88 including *%table1* (24), *%logistic_table* (25) and *%UniLogistic* (26) which have been developed to assist
89 in processing the output from regression procedures, but they are largely limited in terms of flexibility,
90 lack of support for complex survey designs, or are unable to incorporate both categorical and continuous
91 variables in one macro call. For instance, the macro, *%table1*, presents variable names instead of the more
92 meaningful variable labels. The other macros, *%logistic_table*, and *%UniLogistic*, produce output from
93 simple logistic regression but not from multiple logistic regression. Also the *%UniLogistic* macro does
94 not accommodate survey design parameters. Furthermore, these macros lack validation checks for input
95 parameters and also do not export the output into word processing and spreadsheet programs for ease of
96 incorporating into a publication.

97 **Methods**

98 Recognizing the limitations of existing tools, we have designed a SAS macro, *%svy_logistic_regression*,
99 to help overcome these shortcomings while supporting the principles of reproducible research. The macro
100 specifically organizes output from SAS procedures and formats it into quality publication epidemiologic
101 tables containing regression results. We describe the macro functionality and provide an example analysis
102 of a publicly-available dataset and provide access to the source code for the macro to allow others to use
103 and extend it to support their own reproducible research.

104 Our developed SAS macro allows for both simple and multiple logistic regression analysis. Moreover,
105 this SAS macro combines the results from simple and multiple logistic regression analysis into a single

106 quality publication-ready table. The layout of the resulting table is consistent with how models are often
107 presented in the epidemiological and biomedical literature, with unadjusted and adjusted model results
108 presented side by side.

109 The macro, written in SAS software version 9.3 (27), runs logistic regression analysis in a sequential and
110 interactive manner starting with simple logistic regression models followed by multiple logistic regression
111 models using SAS PROC SURVEYLOGISTIC procedure. Frequencies and totals are obtained using
112 PROC SURVEYMEANS and PROC SURVEYFREQ procedures. The final output is then processed
113 using PROC TEMPLATE, PROC REPORT procedures and the output delivery system (ODS).

114 The macro is made up of six sub-macros. The first sub-macro, *%svy_unilogit*, fine-tunes the dataset by
115 applying the conditional statements, and computing the analysis domain size, thus preparing a final
116 analysis dataset. It also prepares the class variables and associated reference categories. It calls the second
117 and third sub-macros, *%svy_logitc* and *%svy_logitn*, to perform separate simple (survey) logistic
118 regression model on each categorical or continuous predictor variable respectively. It further processes
119 results outputs into one table. The fourth sub-macro, *%svy_multilogit*, performs multiple (survey) logistic
120 regression on selected categorical and continuous predictor variables and processes result outputs into one
121 table. The fifth sub-macro, *%svy_printlogit*, combines results from *%svy_unilogit* and *%svy_multilogit*
122 sub-macros and processes the output into an easy to review table which is exported into Microsoft word
123 processing and excel spreadsheet programs. In addition, where survey design variables have been
124 specified the macro automatically incorporates them into the computation. The sixth sub-macro,
125 *%runquit*, is executed after each SAS procedure or DATA step, to enforce in-build SAS validation checks
126 on the input parameters. These include but not limited to checking if the specified dataset exists, ensuring
127 required variables are specified, and verifying that values for reference categories for outcome, domain

128 and categorical variables exist and are valid, as well as checking for logical errors. Once an error is
129 encountered, the macro stops further execution and prints the error message on the log for the user to
130 address it.

131 The macro is generic in that it can be used to analyze any dataset intended to fit a logistic regression model
132 from survey or non-survey settings. It accepts both categorical and continuous predictor variables. Where
133 survey data are used, it allows one to specify design-specific variables such as strata, clusters or weights.
134 Domain analysis for sub-population estimation is also provided for by the macro. Ignoring domain
135 analysis and instead performing a subset analysis will lead to incorrect standard errors. For non-survey
136 settings, the survey input parameters like weights and cluster are set to a default value of 1.

137 The macro also allows the user to explicitly specify the level or category of the binary outcome variable
138 to model as well as reference categories for categorical predictor variables. Further, it runs sequentially
139 by first producing results from simple logistic regression from which the user can select predictor variables
140 to include into the multiple logistic regression, then combine the results of multiple models into a single
141 table. Apart from including only significant predictor variables, based on global/type3 p-values, the user
142 can also choose to include any other variables deemed important by subject matter experts. This flexibility
143 allows for specification of such variables as confounders or effect modifiers even when they are not
144 statistically significant in the simple logistic model. The final output is then processed into a quality
145 publication-ready table and exported into word processing and spreadsheet programs for use in the
146 publication, or if needed, for further hand editing by the authors.

147 The user must provide input parameters which are specified in Table 1. Unless stated (optional), the other
148 parameters must be provided so that the macro can execute successfully. The *outevent* parameter and
149 reference categories for class variables are case sensitive and must be specified in the case they appear in

150 the data dictionary. All other parameters are mainly dataset variables and may be specified in any case.
151 We use lower case for this demonstration. Validation checks enforce these requirements, simplifying
152 debugging errors in macro invocation. The statistician only interacts with sub-macros 1, 4 and 5 by
153 providing input parameters. If a permanent SAS dataset is to be analyzed, the LIBNAME statement can
154 be used to indicate the path or folder where the dataset is located.

155 **Table 1: Input parameters for %svy_unilogit, %svy_multilogit and %svy_printlogit macros**

parameter	Description
%svy_unilogit and %svy_multilogit macros	
dataset	name of input dataset
outcome	name of dependent binary variable of interest e.g., hiv_status
outevent	value label of outcome variable (without quotation) to model e.g., Positive, in the case of modeling Hepatitis A risk factors
catvars	list of categorical variables (nominal or ordinal) separated by space
class	class statement for categorical predictor variables specifying the baseline (reference) category
contvars	list of continuous variables separated by space
condition	(optional) any conditional statements to create and or fine-tune the final analysis dataset specified using one IF statement
strata	(optional) survey stratification variable
cluster	(optional) survey clustering variable
weight	(optional) survey weighting variable
domain	(optional) domain variable for sub-population analysis
print	variable for displaying/suppressing the output table on the output window which takes the values (NO=suppress output, YES=show output)
%svysvy_printlogit macro	
tablename	short name of output table
tabletitle	title of output table
outcome & outevent	same as defined in %svy_unilogit and %svy_multilogit macros
outdir	directory for saving output files
%runquit macro	
syserr	SAS in-build macro variable that checks presence of any system errors

157 **Results**

158 **Example: Analysis of NHANES dataset**

159 We demonstrate the use of our macro in the analysis of the 2013-2014 National Health and Nutrition
160 Examination Survey (NHANES), a suite of complex surveys designed to assess the health and nutritional
161 status of adults and children in the United States (U.S.). In brief, the main objectives of the survey were
162 to estimate and monitor trends in prevalence of selected diseases, risk behaviors and environmental
163 exposures among targeted populations, to explore emerging public health issues, and to provide baseline
164 health characteristics for other administrative use (28).

165 NHANES used a four-stage, stratified sampling design, where counties were selected as primary sampling
166 units (PSUs) using probability proportionate to size (PPS) in the first stage. The second stage involved
167 selecting sections of counties that consisted of a block containing a cluster of households with
168 approximately equal sample sizes per PSU. Dwelling units including households were then selected in the
169 third stage with approximately equal selection probabilities. Individuals within a household were selected
170 in the fourth stage. Stratification was done based on selected demographics characteristics of PSUs.
171 Survey weights were then computed using the various sampling probabilities to account for the complex
172 survey design. The data files are freely available to the public on the NHANES website at:
173 <https://www.cdc.gov/nchs/nhanes/Index.htm>. See (29) for more details regarding the NHANES survey
174 design and contents.

175 The dataset (`clean_nhanes`) used in this example includes participants' socio-demographic
176 characteristics including `riagendr` (Gender), `ridageyr` (Age in years at screening), `ridreth1`,
177 (Race/Hispanic origin), `dmqadfc` (Service in a foreign country), `dmdeduc2` (Education level among

178 adults aged 20+ years), and `dmdmart1` (Marital status). The binary outcome variable is `lhxha`
179 (Hepatitis A antibody test result). The aim of the analysis is to investigate factors associated with a positive
180 test for Hepatitis A antibody among participants aged 20+ years who have served active duty in the U.S.
181 Armed Forces (`dmqmiliz`). Appropriate survey weights `wtmec2yr` (sample weights for participants
182 with a medical examination) were applied. The macros were run with user-defined parameters. The user
183 should explicitly specify the reference category for factor variables and for the binary outcome, as shown
184 in Figure 1.

185 **Figure 1: Sample `%svy_logistic_regression` macro call**

```
%let dir = C:/NHANES/SAS;

* call svy_logistic_regression macro;
option mlogic mprint symbolgen;

* initialize data and outcome variable;
%let dataset = clean_nhanes;
%let outcome = lhxha;
%let outevent = Positive;

* define simple logistic regression model input parameters;
%let classvarb = riagendr(ref="Male") ridageyr<math>cat2</math> (ref=">= 60") ridreth1
(ref="Non-Hispanic White") dmquadfc (ref="No") dmddeduc2 (ref="Some college
or AA degree") dmdmart1 (ref="Divorced");

%let catvarsb = riagendr ridageyr<math>cat2</math> ridreth1 dmquadfc dmddeduc2 dmdmart1;
%let contvarsb = ridageyr;

* fit simple logistic regression model;
%svy_unilogit(dataset = &dataset.,
              outcome = &outcome.,
              outevent = &outevent.,
              catvars = &catvarsb.,
              contvars = &contvarsb.,
              class = &classvarb.,
              weight = wtmec2yr,
              cluster = sdmvpsu,
              strata = sdmvstra,
              domain = dmqmiliz,
              domvalue = 1,
              condition = if ridageyr>=20,
              print = YES);
```

```
* define parameters for selected predictor variables;
%let classvarm = riagendr(ref="Male") ridageyrct2 (ref=">= 60") ridreth1
(ref="Non-Hispanic White") dmqadfc (ref="No") dmdmartl (ref="Divorced");
%let catvarsm = riagendr ridageyrct2 ridreth1 dmqadfc dmdmartl;
%let contvarsm =;

* fit multiple logistic regression model;
%svy_multilogit (dataset = &dataset.,
                outcome = &outcome.,
                outevent = &outevent.,
                catvars = &catvarsm.,
                contvars = &contvarsm.,
                class = &classvarm.,
                weight = wtmec2yr,
                cluster = sdmvpsu,
                strata = sdmvstra,
                domain = dmqmiliz,
                domvalue = 1,
                condition = if ridageyr>=20,
                print = YES);

* output final table;
%svy_printlogit (tablename = logit_table,
                outcome = &outcome.,
                outevent = &outevent.,
                outdir = &dir./output/tables,
                tablettitle = Table 2: Factors associated with Hepatitis A
prevalence among participants who served in the US Armed Forces - NHANES
2013-2014);
```

186

187 The complete SAS output consists of several tables, the majority of which are auxiliary and are used to
188 help in processing the output. Two important output tables are the simple and the multiple logistic
189 regression tables. The simple logistic regression table shows result of bivariate regression as shown in
190 Table 2.

191

192 **Table 2: Output of simple logistic regression model results from %svy_unilogit macro**

Factor	N [@]	Freq ^{&}	OR_CI [§]	p_value ^α	g_p_value ^β
Gender					
Male	473	205 (37.7)	ref		
Female	35	15 (39.7)	1.1 (0.4-2.7)	0.86	0.86
Total	508	220 (37.9)			
Age in years at screening					
>= 60	322	131 (30.7)	ref		
20-39	51	39 (83.0)	11.0 (5.4-22.2)	<.01	<.01
40-59	135	50 (31.2)	1.0 (0.6-1.7)	0.93	
Total	508	220 (37.9)			
Race/Hispanic origin					
Non-Hispanic White	307	114 (34.1)	ref		
Mexican American	23	14 (67.1)	3.9 (1.3-12.3)	0.02	<.001
Non-Hispanic Black	126	59 (46.1)	1.7 (1.2-2.4)	0.01	
Other Hispanic	26	17 (64.3)	3.5 (1.0-11.8)	0.05	
Other Race	26	16 (52.3)	2.1 (0.7-6.8)	0.21	
Total	508	220 (37.9)			
Served in a foreign country					
No	243	86 (27.4)	ref		
Yes	264	134 (48.6)	2.5 (1.4-4.5)	<.01	<.01
Total	507	220 (38.1)			
Education level					
Some college or AA degree	193	92 (41.8)	ref		
9-11th grade	37	16 (33.4)	0.7 (0.4-1.3)	0.27	0.30
College graduate or above	147	51 (32.0)	0.7 (0.4-1.1)	0.09	
High school graduate	122	92 (47.7)	1.0 (0.6-1.5)	0.88	
Less than 9th grade	9	5 (42.6)	1.0 (0.2-4.5)	0.97	
Total	508	220 (37.9)			
Marital status					
Divorced	75	30 (29.4)	ref		
Living with partner	17	9 (60.2)	3.6 (1.1-11.7)	0.03	0.03
Married	311	130 (36.3)	1.4 (0.8-2.3)	0.23	
Never married	48	21 (51.4)	2.5 (1.1-6.1)	0.04	
Separated	11	5 (27.3)	0.9 (0.3-2.6)	0.85	
Widowed	46	25 (44.0)	1.9 (0.9-4.1)	0.11	
Total	508	220 (37.9)			
Age in years at screening	508	220 (37.9)	1.0 (1.0-1.0)	<.01	<.01

193 [@] = Total number of observations

194 [&] = Frequency of sample cases (and weighted row percentages)

195 [§] = Weighted Odds Ratio (95% confidence interval)

196 ^α = Class level p-value

197 $\beta = \text{Global/Type 3 p-value}$

198 The table consists of six variables, namely: `Factor` (risk factor variable), `N` (total frequency of
199 observations), `Freq` (frequency of sample cases and corresponding weighted row percentages), `OR_CI`
200 (weighted odds ratio and 95% confidence interval) `p_value` (class level p-value), `g_p_value`
201 (global/type3 p-value). Typically the analyst/researcher selects statistically-important risk factors based
202 on the global/type3 p-values. From this example, all risk factors except gender and education level were
203 statistically significant. However, based on epidemiological considerations, gender and age are often
204 treated as potential confounder variables. Thus they are included in the multiple logistic regression model
205 regardless of statistical significance. Another important aspect to pick from Table 2 is the frequency
206 columns which show the sample size for each factor and each level of the factor. In this example the
207 expected total measurements for each factor was n=508 out of which 220 (37.8%) tested positive for
208 Hepatitis A antibody. All other factors except service in a foreign country (n=507) had complete
209 information available. The importance of this is to ensure that factors with substantive proportion of
210 complete information are selected for inclusion in the multiple logistic regression model. In addition, the
211 row percentages provide guidance on the choice of reference category of factor variables. However, for
212 ordinal factors it is often advisable to use the lowest or highest category as reference, depending on the
213 outcome of interest. After selecting all important variables, the `%svy_multilogit` macro is then executed.
214 The `%svy_printlogit` macro automatically processes the output into a quality easy to review output as
215 shown in Table 3.

216

217 **Table 3: Quality publication-ready output from the %svy_printlogit macro combining results from**
 218 **%svy_unilogit and %svy_multilogit macros**

Characteristic	Hepatitis A antibody [€]		Unadjusted odds ratios			Adjusted odds ratios		
	Total [¥]	Positive [£] n (%) ^{&}	OR ^ξ (95% CI) ^{\$}	p-value ^α	Type3 ^β p-value	OR (95% CI)	p-value	Type3 p-value
Gender								
Male	473	205 (43.3)	ref					
Female	35	15 (42.9)	1.1 (0.4-2.7)	0.86	0.86	1.0 (0.3-3.7)	1.00	1.00
Total	508	220 (43.3)						
Age in years at screening								
>= 60	322	131 (40.7)	ref					
20-39	51	39 (76.5)	11.0 (5.4-22.2)	<.01	<.01	13.8 (5.6-34.0)	<.01	<.01
40-59	135	50 (37)	1.0 (0.6-1.7)	0.93		1.3 (0.6-2.8)	0.54	
Total	508	220 (43.3)						
Race/Hispanic origin								
Non-Hispanic White	307	114 (37.1)	ref					
Mexican American	23	14 (60.9)	3.9 (1.3-12.3)	0.02	<.01	3.4 (1.1-10.4)	0.03	0.01
Non-Hispanic Black	126	59 (46.8)	1.7 (1.2-2.4)	0.01		1.9 (1.1-3.2)	0.02	
Other Hispanic	26	17 (65.4)	3.5 (1.0-11.8)	0.05		3.4 (0.6-18.6)	0.16	
Other Race	26	16 (61.5)	2.1 (0.7-6.8)	0.21		1.4 (0.3-6.9)	0.65	
Total	508	220 (43.3)						
Served in a foreign country								
No	243	86 (35.4)	ref					
Yes	264	134 (50.8)	2.5 (1.4-4.5)	<.01	<.01	3.1 (1.9-5.0)	<.01	<.01
Total	507	220 (43.4)						
Education level								
Some college or AA degree	193	92 (47.7)	ref					
9-11th grade	37	16 (43.2)	0.7 (0.4-1.3)	0.27	0.30			
College graduate or above	147	51 (34.7)	0.7 (0.4-1.1)	0.09				
High school graduate	122	56 (45.9)	1.0 (0.6-1.5)	0.88				
Less than 9th grade	9	5 (55.6)	1.0 (0.2-4.5)	0.97				
Total	508	220 (43.3)						
Marital status								
Divorced	75	30 (40.0)	ref					
Living with partner	17	9 (52.9)	3.6 (1.1-11.7)	0.03	0.03	1.8 (0.6-5.1)	0.27	0.02
Married	311	130 (41.8)	1.4 (0.8-2.3)	0.23		1.6 (1.0-2.4)	0.06	

Never married	48	21 (43.8)	2.5 (1.1-6.1)	0.04		1.5 (0.9-2.7)	0.13	
Separated	11	5 (45.5)	0.9 (0.3-2.6)	0.85		0.9 (0.3-2.9)	0.88	
Widowed	46	25 (54.3)	1.9 (0.9-4.1)	0.11		3.2 (1.5-6.7)	<.01	
Total	508	220 (43.3)						
Age in years at screening	508	220 (43.3)	1.0 (1.0-1.0)	<.01	<.01			

219 ¥ = Total number of observations

220 € = Outcome variable label

221 £ = Outcome value label of category of interest

222 & = Frequency of sample cases (and weighted row percentages)

223 ξ = Weighted Odds Ratio

224 \$ = Weighted 95% confidence interval for Odds Ratio

225 α = Class level p-value

226 β = Global/Type3 p-value

227

228 **Discussion**

229 This paper presents an elegant and flexible SAS macro, *%svy_logistic_regression*, for producing quality
230 publication-ready tables from unadjusted and adjusted logistic regression analyses. Even though a number
231 of SAS macros are available on the internet for processing output from logistic regression into a
232 publication-ready table, they are complex to follow and/or have limited features, thus restricting their
233 adoption. Many macros are not generic and hence can only be used with the data for which they were
234 designed.

235 The SAS macro presented here is generalized, highly suitable to handle different scenarios, and is simple
236 to implement and invoke from user macros. In addition, our macro includes the row or column total and
237 frequency of prevalent cases of each variable level, which can immediately allow the analyst/researcher
238 to identify levels with sparse data. Row percentages help the researcher in the choice of reference category.
239 Global or type3 p-values shows whether or not a variable is an important predictor. Individual p-values
240 shows if a given variable category is comparable to the reference category. The macro provides validation
241 checks on the input parameters including the dataset, variables and values of variables to ensure that the
242 analyst obtains valid estimates. The output of this SAS macro helps improve efficiency of knowledge
243 generation by reducing the steps required from analysis to clear and concise presentation of results.

244 **Conclusion**

245 As our contribution to the emerging field of reproducible research, we have provided source code for the
246 SAS macro as well as expected outputs using a publicly available dataset. By publishing this macro, it
247 will allow other SAS macro programmers and users to verify and build upon this code. Production of

248 publication-quality tables is increasingly important as data analyses become more complex, involving
249 larger datasets and requiring more sophisticated computations and tabulation, notwithstanding the need
250 for quick results. This macro helps to make data analysis results readily available, and allows one to
251 publish data summaries in a single document, thus allowing others to easily execute the same code to
252 obtain the same results. The quality, publication-ready results from this macro are suitable for direct
253 inclusion in manuscripts for peer-reviewed journals. The macro can also be used to routinely generate
254 standardized tables. This is especially useful for disease surveillance systems where the same analyses are
255 repeated on a quarterly or annual basis. We hope the published results from this macro will provide

256 **Supporting information**

257 **Supporting results dataset**

258 The NHANES dataset supporting the conclusions of this article is freely available to the public on
259 the NHANES website at: <https://www.cdc.gov/nchs/nhanes/Index.htm>.

260 **Supporting software**

261 The source code for this macro is available online at <https://github.com/kmuthusi/generic-sas-macros> for
262 public access and has been licensed under the terms of the Apache Software License and therefore is
263 licensed under ASL v2 or later. A copy on this license is available at
264 <http://www.apache.org/licenses/LICENSE-2.0.html>. Sample SAS program call for the macro named
265 “*svy logistic regression anafile.sas*” is provided.

266 **Funding**

267 This work was supported by the President’s Emergency Plan for AIDS Relief (PEPFAR) through the U.S.
268 Centers for Disease Control and Prevention (CDC). The funders had no role in study design, data
269 collection and analysis, decision to publish, or preparation of the manuscript.

270 **Acknowledgement**

271 We thank our colleagues in the Surveillance and Epidemiology branch for testing the SAS macro and
272 providing valuable feedback for improvement and also for thoroughly reviewing the manuscript.

273 **References**

- 274 1. Peng RD, Dominici F, Zeger SL. Reproducible epidemiologic research. American Journal of
275 Epidemiology. 2006;163(9):783-9.
- 276 2. Peng RD. Reproducible research in computational science. Science. 2011;334(6060):1226-7.
- 277 3. Peng RD. Reproducible research and Biostatistics. Biostatistics. 2009;10(3):405-8.
- 278 4. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JP. Reproducible Research Practices and
279 Transparency across the Biomedical Literature. PLoS Biol. 2016;14(1):e1002333.
- 280 5. Arnold Tim, Kuhfeld Warren F. Using SAS and LATEX to Create Documents with Reproducible
281 Results. URL: <http://supportsas.com/resources/papers/proceedings12/324-2012pdf>. 2012.
- 282 6. Wellcome Trust. Policy on data, software and materials management and sharing 2017 [
283 7. Wellcome Trust. Open access policy 2017 [
284 8. U S Department of Health and Human Services. Open Government Plan. 2016. p. 48-9.
- 285 9. Leek JT, Peng RD. Opinion: Reproducible research can still be wrong: adopting a prevention
286 approach. Proc Natl Acad Sci U S A. 2015;112(6):1645-6.
- 287 10. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical
288 journals. British Medical Journal (Clinical research ed). 1983;286(6376):1489-93.
- 289 11. Lang TA, Altman DG. Basic statistical reporting for articles published in clinical medical journals:
290 the SAMPL Guidelines. In: Smart P, Maisonneuve H, Polderman A (eds). Science Editors' Handbook,
291 European Association of Science Editors. 2013.

- 292 12. Nala R, Cummings B, Horth R, Inguane C, Benedetti M, Chissano M, et al. Men who have sex
293 with men in Mozambique: identifying a hidden population at high-risk for HIV. *AIDS Behavior*.
294 2015;19(2):393-404.
- 295 13. Moore DM, Cui Z, Lachowsky N, Raymond HF, Roth E, Rich A, et al. HIV Community Viral Load
296 and Factors Associated With Elevated Viremia Among a Community-Based Sample of Men Who Have
297 Sex With Men in Vancouver, Canada. *Journal of Acquired Immune Deficiency Syndrome*.
298 2016;72(1):87-95.
- 299 14. Cherutich P, Kim AA, Kellogg TA, Sherr K, Waruru A, De Cock KM, et al. Detectable HIV Viral
300 Load in Kenya: Data from a Population-Based Survey. *PLoS One*. 2016;11(5):e0154318.
- 301 15. Oluoch T, Katana A, Kwaro D, Santas X, Langat P, Mwalili S, et al. Effect of a clinical decision
302 support system on early action on immunological treatment failure in patients with HIV in Kenya: a
303 cluster randomised controlled trial. *The Lancet HIV*. 2016;3(2):e76-e84.
- 304 16. Jann B. Making regression tables from stored estimates. *The Stata Journal*, URL:
305 <http://wwwstata-journal.com/articlehtml?article=st0085>. 2005;5(3):288-308.
- 306 17. Jann B. Making regression tables simplified. *The Stata Journal*, URL: [http://wwwstata-](http://wwwstata-journal.com/articlehtml?article=st0085_1)
307 [journal.com/articlehtml?article=st0085_1](http://wwwstata-journal.com/articlehtml?article=st0085_1). 2007;7(2):227-44.
- 308 18. Brady T. REFORMAT: Stata module to reformat regression output. *Statistical Software*
309 *Components S426304*, Boston College Department of Economics, revised 06 Oct 2002, URL:
310 <https://ideasrepecorg/c/boc/bocode/s426304html>. 2002.
- 311 19. Wada R. OUTREG2: Stata module to arrange regression outputs into an illustrative table.
312 *Statistical Software Components S456416*, Boston College Department of Economics, revised 17 Aug
313 2014 URL: <https://ideasrepecorg/c/boc/bocode/s456416html>. 2005.

- 314 20. Hlavac M. stargazer: Well-Formatted Regression and Summary Statistics Tables. R package
315 version 5.2.1. <https://CRAN.R-project.org/package=stargazer>. 2018.
- 316 21. Robinson D, Gomez M, Demeshev B, Menne D, Nutter B, Luke J, et al. broom: Convert Statistical
317 Analysis Objects into Tidy Data Frames. URL: <https://cran.r-project.org/package=broom>. 2017.
- 318 22. Gohel D, Nazarov M. flextable: Functions for Tabular Reporting. URL: [https://cran.r-](https://cran.r-project.org/package=flextable)
319 [project.org/package=flextable](https://cran.r-project.org/package=flextable). 2018.
- 320 23. SAS Institute Inc. SAS/STAT® 9.3 User's Guide. Cary, NC: SAS Institute Inc; 2011.
- 321 24. Gravely A, Clothier B, Nugent S. Creating an Easy to Use, Dynamic, Flexible Summary Table
322 Macro with P-Values in SAS® for Research Studies. Proceedings from MidWest SAS Users Group Paper
323 AA072014.
- 324 25. Qi J. Automating the Process of Generating Publication Quality Regression Tables through SAS®
325 Base Programming. Proceedings from MiwWest SAS Users Group Paper BB232016.
- 326 26. Dhand NK. UniLogistic: A SAS Macro for Descriptive and Univariable Logistic Regression
327 Analyses. Journal of Statistical Software. 2010;35(1):1-15.
- 328 27. SAS Institute Inc. Base SAS® 9.3. Cary, NC: SAS Institute Inc; 2011.
- 329 28. Centers for Disease Control and Prevention. Centers for Disease Control and Prevention (CDC).
330 National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data.
331 Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and
332 Prevention, 2013-2014, URL: <https://www.cdc.gov/nchs/nhanes/Index.htm>: National Center for Health
333 Statistics (NCHS); 2015 [

- 334 29. Johnson CL, Dohrmann SM, Burt VL, Mohadjer LK. National Health and Nutrition Examination
335 Survey: Sample design, 2011–2014. National Center for Health Statistics. Vital and Health Statistics.
336 2014;2(162).