Paper 173-2007

A Purposeful Selection of Variables Macro for Logistic Regression

Zoran Bursac[1], C. Heath Gauss[1], D. Keith Williams[1], and David Hosmer[2]
University of Arkansas for Medical Sciences[1] and University of Massachusetts[2]

ABSTRACT

The main problem in any model-building situation is to choose from a large set of covariates those that should be included in the "best" model. A decision to keep a variable in the model might be based on the clinical or statistical significance. There are several variable selection algorithms embedded in SAS PROC LOGISTIC. Those methods are mechanical and as such carry some limitations. Hosmer and Lemeshow describe a purposeful selection of covariates algorithm within which an analyst makes a variable selection decision at each step of the modeling process. In this paper we introduce a macro, *%PurposefulSelection*, which automates that process. The macro is based on the following algorithm: (1) fit a univariate model with each covariate, (2) select as candidates for a multivariate model those significant at some chosen alpha level, (3) identify those variables that are not significant in the multivariate model at some arbitrary alpha level, (4) fit a reduced model and evaluate confounding by change in parameter estimates, (5) repeat steps 3 and 4 until the model contains significant covariates and/or confounders and (6) add back in the model, one at a time, any variable not originally selected, keep any that are significant, and reduce the model following steps 3 and 4. At the end of step 6, the analyst will have a "main effects model." Performance of the macro is illustrated with the application to the Hosmer and Lemeshow Worchester Heart Attack Study (WHAS) data.

Keywords: logistic regression, SAS PROC LOGISTIC, variable selection algorithm, purposeful selection, confounding

BACKGROUND

The criteria for inclusion of a variable in the model vary between problems and disciplines. The common approach to statistical model building is minimization of variables until the most parsimonious model that describes the data is found which also results in numerical stability and generalizability of the results. Some methodologists suggest inclusion of all clinical and other relevant variables in the model regardless of their significance in order to control for confounding. This approach, however, can lead to numerically unstable estimates and large standard errors. This paper is based on the purposeful selection of variables in logistic regression as proposed by Hosmer and Lemeshow [2000].

Several variable selection methods are available in SAS PROC LOGISTIC. The simplest method (and the default) is SELECTION=NONE, for which PROC LOGISTIC fits the complete model as specified in the MODEL statement. The other commonly used methods and the ones of focus in this paper are FORWARD for forward selection, BACKWARD for backward elimination, and STEPWISE for stepwise selection, [SAS Institute Inc., 2004].

When SELECTION=FORWARD, PROC LOGISTIC computes the score chi-square statistic for each effect not in the model and examines the largest of these statistics. If it is significant at some entry level, the corresponding effect is added to the model. Once an effect is entered in the model, it is never removed from the model. The process is repeated until none of the remaining effects meet the specified level for entry [SAS Institute Inc., 2004].

When SELECTION=BACKWARD, results of the Wald test for individual parameters are examined. The least significant effect that does not meet the level for staying in the model is removed. Once an effect is removed from the model, it remains excluded. The process is repeated until no other effect in the model meets the specified level for removal. [SAS Institute Inc., 2004].

The SELECTION=STEPWISE option is similar to the SELECTION=FORWARD option except that effects already in the model do not necessarily remain. Effects are entered into and removed from the model in such a way that each forward selection step may be followed by one or more backward elimination steps. The stepwise selection process terminates if no further effect can be added to the model or if the effect just

entered into the model is the only effect removed in the subsequent backward elimination [SAS Institute Inc., 2004].

At the 2003 North East SAS Users Group meeting (NESUG), Hegewald, Pfahlberg, and Uter [2003] introduced a backward manual selection (BMS) macro for logistic regression. The BMS macro follows the variable selection algorithm proposed by Kleinbaum, Kupper, and Morgenstein [1982] by first eliminating interactions based on their significance level followed by selection or elimination of potential confounders if they do not influence the estimated odds ratio (OR) of the main exposure of interest and/or have a negative effect on the global model-fit.

The purposeful selection algorithm follows a slightly different logic as proposed by Hosmer and Lemeshow [2000]. The selection process begins by a univariate analysis of each variable. Any variable having a significant univariate test at some arbitrary level is selected as a candidate for the multivariate analysis. We base this on the Wald test from logistic regression and p-value cut-off point of 0.25. More traditional levels such as 0.05 can fail in identifying variables known to be important. In the iterative process of variable selection, covariates are removed from the model if they are non-significant and not a confounder. Significance is evaluated at the 0.1 alpha level and confounding as a change in any parameter estimate greater than 20%. The macro allows the user to specify all decision criteria. At the end of this iterative process, the model contains significant covariates and confounders. At this point any variable not selected for the original multivariate model is added back one at a time, with significant covariates and confounders retained earlier. Any that are significant at the 0.1 level are put in the model, and the model is iteratively reduced as before but only for the variables that were additionally added. At the end of this final step, the analyst is left with the preliminary main effects model. Unlike the BMS macro, the evaluation of interactions is not currently embedded in the purposeful selection macro.

METHODS

The main *%PurposefulSelection* macro consists of three calls to sub-macros, *%ScanVar*, *%UniFit* and *%MVFit*. The *%ScanVar* sub-macro scans the submitted covariates and prepares them for the univariate analysis. The *%UniFit* sub-macro fits all univariate models and creates a data set with the candidate variables for the multivariate analysis. The *%MVFit* sub-macro iteratively fits multivariate models while evaluating the significance and confounding effect of each candidate variable as well as those that were not originally selected. A flowchart of the macro is presented in Figure 1.

Table 1. Macro variables.

DATASET	Input data set
OUTCOME	Main outcome (y)
COVARIATES	All covariates ($x_1 \ldots x_j$)
PVALUEI	Inclusion criteria for multivariate model
PVALUER	Retention criteria for multivariate model
CHBETA	% change in parameter estimate indicating confounding

User must define several macro variables as shown in Table 1. Macro variable *DATASET* corresponds to the data set to be analyzed. Macro variable *OUTCOME* is the main outcome of interest and should be a binary variable (also known as the dependent variable). Macro variable *COVARIATES* represents a set of predictor variables which can all be continuous, binary, or a mix of the two. In the case of a polytomous covariate, dummy variables must be created before invoking the macro and specified as separate variables. All covariates specified here are assumed to be of equal importance. Macro variable *PVALUEI* defines the alpha level for the univariate model at which any of the covariates will be considered as a candidate for the multivariable analysis. Macro variable *PVALUER* defines the retention criteria for the multivariate model at which the variable will remain in the model. Finally, the macro variable *CHBETA* represents the percent change in a parameter estimate (beta) above which a covariate that is removed from the model as non-significant will be considered a confounder and placed back in the model. Even though we recommend inclusion and retention criteria to be set at 0.25 and 0.1 respectively, and confounding at 20% change, these parameters can be directly controlled by the analyst, since they are coded as macro variables.

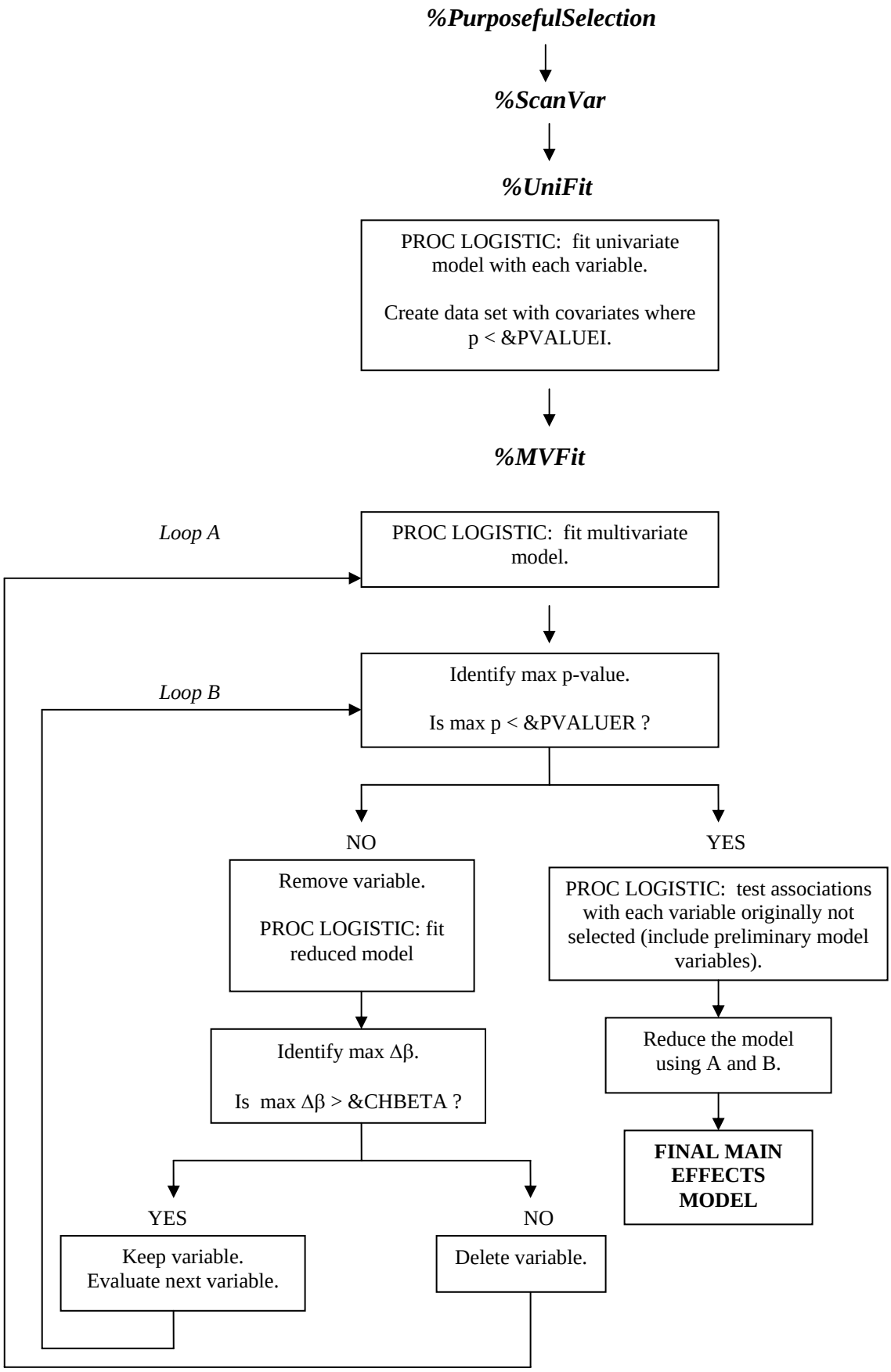

Figure 1. *%PurposefulSelection* macro flow chart.

WHAS DATA EXAMPLE

A subset of observations (N=307) and variables from the WHAS data published by Hosmer and Lemeshow [1999] was used to compare the results of variable selections between the *%PurposefulSelection* macro and each of the three methods available in PROC LOGISTIC as described above.

Table 2. WHAS data variables.

FSTAT	Status as of last follow-up (1=Dead, 0=Alive)
AGE	Age at hospital admission (Years)
SEX	Gender (0=Male, 1=Female)
HR	Initial heart rate (Beats per minute)
BMI	Body mass index (kg/m^2)
CVD	History of cardiovascular disease (0=No, 1=Yes)
AFB	Atrial fibrillation (0=No, 1=Yes)
SHO	Cardiogenic shock (0=No, 1=Yes)
CHF	Congestive heart complications (0=No, 1=Yes)
AV3	Complete heart block (0=No, 1=Yes)
MIORD	MI order (0=First, 1=Recurrent)
MITYPE	MI type (0=non - Q-wave, 1=Q-wave)

The main outcome of interest was vital status at the last follow-up, dead (FSTAT=1) versus alive (FSTAT=0). The eleven covariates listed in Table 2 were treated as equally important. The following macro call was used to invoke purposeful selection of variables from the WHAS data set:

%***PurposefulSelection*** (whas, fstat, age sex hr bmi cvd afb sho chf av3 miord mitype, **0.25**, **0.1**, **20**);

Table 3 shows the results of variable retention from our macro and PROC LOGISTIC selection procedures. The *%UniFit* sub-macro identified 9 covariates initially as potential candidates for the multivariate model at the 0.25 alpha level based on the Wald chi-square statistic. Those included AGE, SEX, HR, BMI, CVD, AFB, SHO, CHF, and MIORD. In the *%MVFit* sub-macro, four of them (SEX, CVD, AFB, and CHF) were eliminated one at a time because they were not significant in the multivariate model at the alpha level of 0.1 and when taken out, did not change any remaining parameter estimates by more than 20%. The variable BMI was also not significant at the 0.1 alpha level but changed the parameter estimate for the MIORD covariate by more than 20% when taken out; therefore, it remained in the model as a confounder. The maximum p-value of the remaining variables AGE, SHO, HR and MIORD was less than 0.1 at which point the variables originally set aside were considered.

Table 3. Variables retained in final models.

Purposeful Selection	*p-value*	*Forward, Backward, Stepwise*	*p-value*
AGE	<0.0001	AGE	<0.0001
SHO	0.0018	SHO	0.0039
HR	0.0025	HR	0.0011
MITYPE	0.091	MITYPE	0.0149
MIORD	0.1087	AV3	0.0672
BMI	0.2035		

Out of the remaining two variables set aside initially because they were not significant at the 0.25 level (AV3 and MITYPE), MITYPE made it back in the model when tested (one at a time) with the five retained

covariates because it was significant at the 0.1 alpha level. The addition of MITYPE confounded the relationship between MIORD and FSTAT, hence the change in the MIORD p-value from 0.0324 to 0.1087.

All three selection procedures available in PROC LOGISTIC resulted in the same model (Table 3). Changing the SLENTRY and SLSTAY options could potentially change the final models, but mostly as a consequence of the default alpha values, resulting in the inclusion of non-significant, non-confounding noise variables. While the resulting model contains only significant covariates, it did not retain the confounder BMI or the variable MIORD which were retained by the *%PurposefulSelection* macro. On the other hand, the variable AV3 was retained.

DISCUSSION

The human modeling process still remains the most powerful one. We can attempt to control for as many situations as possible through automated computer algorithms, but that is still not an adequate replacement for a skilled analyst making decisions at each step of the modeling process.

The advantage of the *%PurposefulSelection* macro comes when the analyst is interested in risk factor modeling and not just mere prediction. The algorithm is written in such a way that, in addition to significant covariates, it retains important confounding variables that may or may not be significantly associated with the outcome.

LIMITATIONS

There are a couple of limitations to this algorithm. First, variables not selected initially for the multivariate model are tested later on with the selected set of covariates one at a time. If two or more variables are significant when put in the model jointly, they are going to be missed. However, being significant jointly may indicate multicollinearity, in which case the analyst may choose to use only one of those as a proxy or not at all. Secondly, if two non-significant covariates confound each other, they are going to be retained as confounders since we assume that all covariates are equally important. In a situation where that happens, the analyst should probably consider retaining the two covariates if they are significant at the 0.25 level, indicating some reasonable association with the outcome. Otherwise, the analyst should probably exclude both from the model as meaningless confounders.

CONCLUSION

If an analyst is in need of an algorithm that will help guide the retention of significant covariates as well as confounding ones, this macro will provide that. Analysts should use this macro as a tool that helps with decisions about the final model, not as a definite answer. One should always carefully examine the model provided by this macro and determine why the covariates were retained before proceeding.

REFERENCES

1. Hegewald, J., Pfahlberg, A., and Uter, W. 2003. “A Backwards-Manual Selection Macro for Binary Logistic Regression in the SAS v.8.02 PROC LOGISTIC Procedure”. *Proceedings of the North East SAS Users Group*.
2. Hosmer, D.W., and Lemeshow, S. 2000. *Applied Logistic Regression.* New York: Wiley.
3. Kleinbaum, D., and Kupper, L.L., et al. 1982. *Epidemiological Research, Principles and Quantitative Methods*. New York: Wiley.
4. SAS Institute Inc. 2004. *SAS/STAT User’s Guide, Version 9.1*. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

The *%PurposefulSelection* macro will be provided as requested.

Zoran Bursac
Biostatistics
Fay W. Boozman College of Public Health
University of Arkansas for Medical Sciences
4301 W. Markham, Slot 781
Little Rock, AR 72205
Work Phone: (501) 526-6723
Fax: (501) 526-6729
E-mail: zbursac@uams.edu
Web: www.uams.edu/biostat/bursac/