

Purposeful Selection of Variables in Logistic Regression: Macro and Simulation Results

Zoran Bursac¹, C. Heath Gauss¹, D. Keith Williams¹, David Hosmer²

¹Biostatistics, University of Arkansas for Medical Sciences, Little Rock, AR, 72205

²Biostatistics, University of Massachusetts, Amherst, MA, 01003

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 [2000] 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 this process. We conduct a simulation study to compare the performance of this algorithm with three well documented variable selection procedures in SAS PROC LOGISTIC: FORWARD, BACKWARD, and STEPWISE. Results and implications are discussed in more detail.

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

1. 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 (F), 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 (B), 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 (S) 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 Northeast 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.

2. Methods

The main *%PurposefulSelection* (P) 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 ($X_1 \dots X_j$)
PVALUEI	Inclusion criteria for multivariate model
PVALUER	Retention criteria for multivariate model
CHBETA	% change in parameter estimate indicating confounding
PVALUENC	Inclusion criteria for non-candidate

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 a covariate will be considered as a candidate for the multivariate analysis. Macro variable *PVALUER* defines the retention criteria for the multivariate model at which a variable will remain in the model. 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. Finally, macro variable *PVALUENC* defines the inclusion criteria for non-candidate variables, allowing then to make it back into the model. We recommend this value be set at 0.15 or 0.1.

%PurposefulSelection

%ScanVar

%UniFit

PROC LOGISTIC: fit univariate model with each variable.
Create data set with covariates where $p < \&PVALUEI$.

%MVFit

Loop A

PROC LOGISTIC: fit multivariate model.

Loop B

Identify max p-value.
Is max $p < \&PVALUER$?

NO

YES

Remove variable.
PROC LOGISTIC: fit reduced model

PROC LOGISTIC: test associations with each variable originally not selected (include preliminary model variables).

Identify max $\Delta\beta$.
Is max $\Delta\beta > \&CHBETA$?

Reduce the model using A and B.

YES

NO

Keep variable.
Evaluate next variable.

Delete variable.

FINAL MAIN EFFECTS MODEL

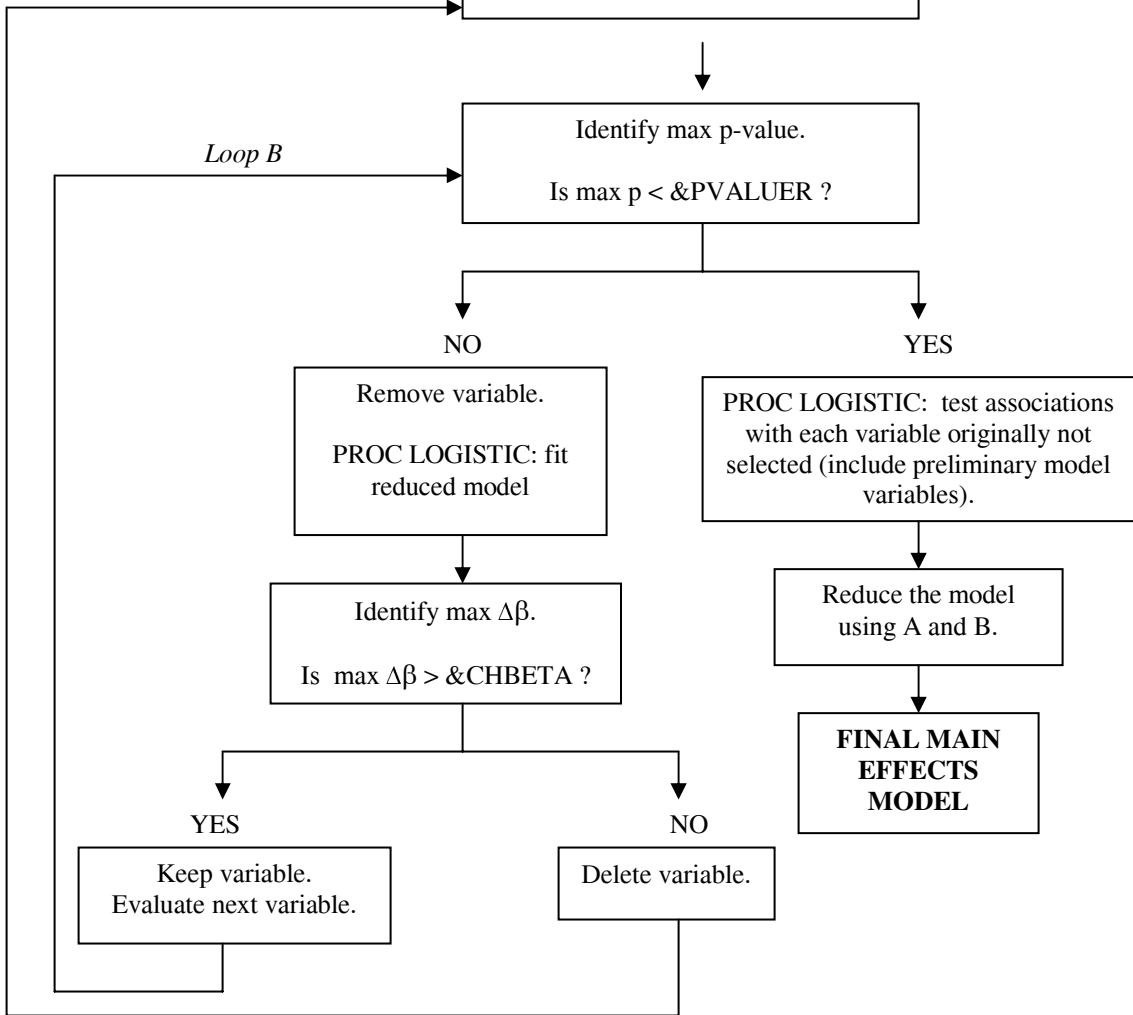


Figure 1. *%PurposefulSelection* macro flow chart.

3. Simulation Study

We conducted two simulation studies to evaluate the performance of the *%PurposefulSelection* macro. In the first simulation, we started with the assumption that we had 6 equally important covariates (X_1, \dots, X_6 such that $X_j \sim U(-6, 6)$ for $j=1, \dots, 6$), three of which were significant and three that were not. We set $\beta_0 = -0.6$, $\beta_1 = \beta_2 = \beta_3 = 0.122$ and $\beta_4 = \beta_5 = \beta_6 = 0$. Therefore the true logit we sampled from was

$$\text{logit} = -0.6 + 0.122X_1 + 0.122X_2 + 0.122X_3 + 0X_4 + 0X_5 + 0X_6.$$

We conducted 1000 simulation runs for each of the 6 conditions in which we varied the sample size (n=60, 120, 240, 360, 480, and 600). The summary measure of the algorithm performance was the percent of times each variable selection procedure retained only X_1 , X_2 , and X_3 as the final model. For P selection, *CHBETA* was set to 20% and *PVALUENC* to 0.1, even though confounding was not simulated in this portion of the study.

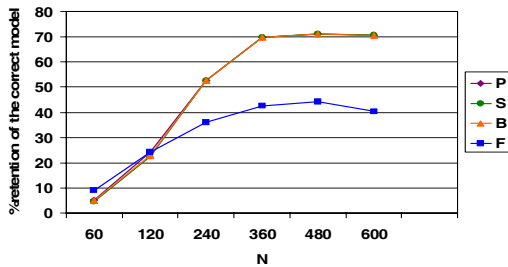


Figure 2. Results of simulation study 1.

Figure 2 shows the proportion of times that the correct model was retained for four selection procedures under various sample sizes. Correct retention increases with sample size, and it is almost identical for P, S and B. F selection does not perform as well as the other three with exception of lower sample size levels.

In the second simulation, we started with the same assumption, that the 6 covariates were equally important, two of which were significant, one that was the confounder, and three that were not significant. We assumed that $X_1 \sim \text{Bernoulli}(0.5)$, the confounder $X_2 \sim U(-6, 3)$ if $X_1 = 1$ and $X_2 \sim U(-3, 6)$ if $X_1 = 0$, and $X_3 - X_6 \sim U(-6, 6)$. We created the confounder X_2 by making the distribution of that variable dependent on X_1 .

We set $\beta_0 = -0.6$, $\beta_1 = 1.2$, $\beta_2 = 0.1$, $\beta_3 = 0.122$, and $\beta_4 = \beta_5 = \beta_6 = 0$. Therefore the true logit we sampled from was

$$\text{logit} = -0.6 + 1.2X_1 + 0.1X_2 + 0.122X_3 + 0X_4 + 0X_5 + 0X_6.$$

We conducted 1000 simulation runs for each of the 24 conditions in which we varied the sample size (n=60, 120, 240, 360, 480, and 600), *CHBETA* (20% and 15%), and *PVALUENC* (0.15 and 0.1). Similarly, the summary measure of the algorithm performance was the percent of times each variable selection procedure retained only X_1 , X_2 , and X_3 in the final model.

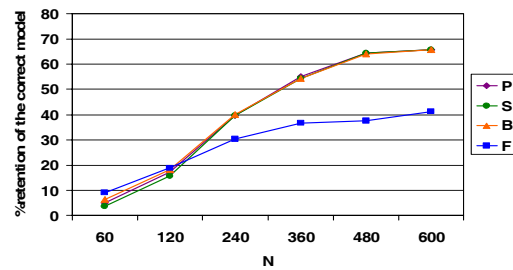


Figure 3. *CHBETA*=20%, *PVALUENC*=0.1 .

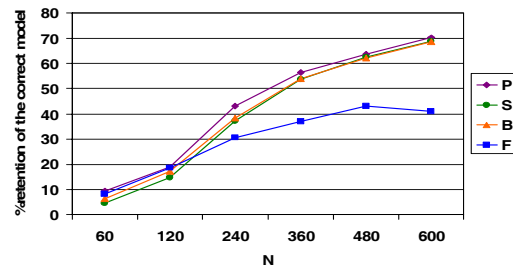


Figure 4. *CHBETA*=20%, *PVALUENC*=0.15 .

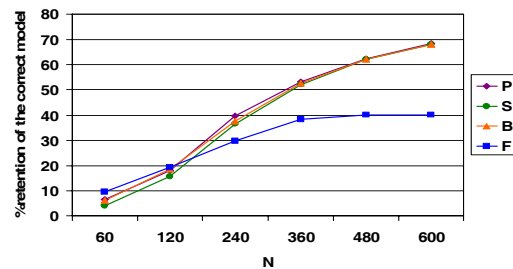


Figure 5. *CHBETA*=15%, *PVALUENC*=0.1 .

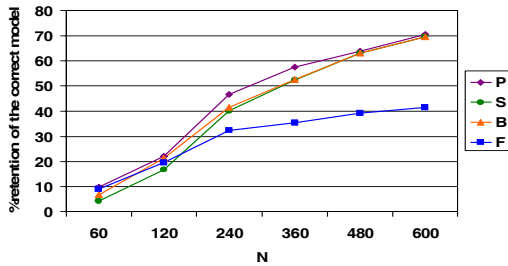


Figure 6. $CHBETA=15\%$, $PVALUENC=0.15$.

Figures 3-6 show the proportion of times that the correct model was retained for four selection procedures under 24 simulated conditions.

Again proportion of correctly retained models increases with sample size for all selection methods. At the lower sample size levels, no procedure performs very well. F selection does the best with exception of the situation when $PVALUENC$ is set to 0.15, where P performs better. With the larger samples like 480 and 600, P, S, and B selections converge toward a close proportion of correct model retention while F selection does notably worse. With confounding present, P selection retains the correct model a larger proportion of times for all six sample sizes when $CHBETA$ is set to either 20% or 15% and $PVALUENC$ to 0.15 as compared to the other three methods. Under other scenarios it retains the correct model slightly more often compared to the others, and mainly for sample sizes in the range 240-360.

In addition to the mentioned simulation conditions, we tampered with the coefficient of the confounding variable X_2 , by making it more significant at 0.13, and less significant at 0.07. We show the results for both scenarios with $CHBETA$ set to 15%, and $PVALUENC$ at 0.15.

When $\beta_2=0.13$, Figure 7 shows that P, B, and as sample size gets larger, S perform comparably, retaining a similar proportion of correct models. This is primarily due to the fact that X_2 becomes significant in larger proportion of simulations and is retained by those procedures because of its significance and not confounding effect. F selection again does worse than the three previously mentioned selection procedures.

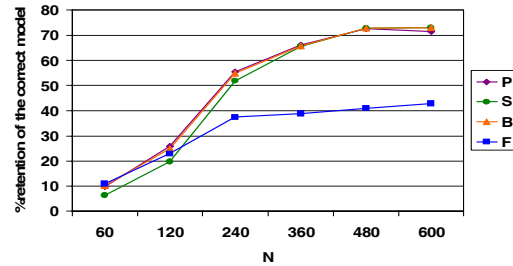


Figure 7. $\beta_2=0.13$, $CHBETA=15\%$, $PVALUENC=0.15$.

When $\beta_2=0.07$, Figure 8 shows that P performs better across all sample sizes than other variable selection procedures; however, the proportion of correctly retained models is lower for all procedures. This is a result of the fact that X_2 becomes non-significant in more simulations and is not retained, or is missed as a result. Figure 8 also shows how X_2 is picked up by P selection due to its confounding effect which is still present.

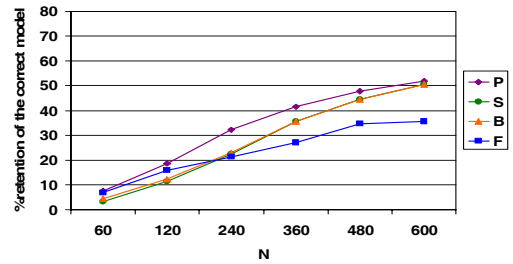


Figure 8. $\beta_2=0.07$, $CHBETA=15\%$, $PVALUENC=0.15$.

4. 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.

The simulation study demonstrates that the purposeful selection algorithm identifies and retains confounders correctly at a larger rate than other selection procedures, particularly in instances where the significance level of a confounder is between 0.1 and 0.15 when the other algorithms would not retain it.

4.1 Limitations

There are some 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. Therefore 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. This algorithm was not designed to force all dummy variables in if one is significant, but the other selection procedures have this limitation as well, unless you force them in with the INCLUDE statement. However, it is not possible to know a priori whether one of the dummy variables will be significant. If one of the dummy variables is retained as significant, the analyst can manually insert the rest of them in the model.

5. Conclusions

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. In order to improve the chances of retaining meaningful confounders, we recommend setting *CHBETA* to 15% and *PVALUENC* to 0.15. 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

- 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*.
- Hosmer, D.W., and Lemeshow, S. 2000. *Applied Logistic Regression*. New York: Wiley.
- Kleinbaum, D., and Kupper, L.L., et al. 1982. *Epidemiological Research, Principles and Quantitative Methods*. New York: Wiley.
- 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/