

Backward Variable Selection for Logistic Regression Based on Percentage Change in Odds Ratio

Evan Kwiatkowski, University of North Carolina at Chapel Hill;
Hannah Crooke, PAREXEL International and University of North Carolina at Charlotte;
Kathy Roggenkamp, University of North Carolina at Chapel Hill

ABSTRACT

Variable selection is a fundamental component of statistical modeling. A common variable selection method used in health sciences is backward variable selection, which iteratively removes variables based on their relevance to the model. Often, automated backward variable selection procedures determine variable relevance based on overall statistical significance. However, many epidemiologists, including formative thinkers Greenland and Robins, favor a "change-in-estimate" approach to variable selection rather than an overall significance approach. We developed a SAS® software macro to implement a backward variable selection procedure for logistic regression using the "change-in-estimate" method.

Our macro implements backwards variable selection in the logistic regression model in the situation where there is a single independent variable (IV) and single dependent variable (DV) of interest, with additional covariates that are eligible for removal based on their relevance to the model. This relevance is based on the percentage change in odds ratio between the IV and DV in a full model including all additional covariates and a reduced model which removes a single covariate at a time.

This macro provides epidemiologists and other health science professionals with a theoretically sound option for automated backward variable selection in logistic regression, and is an extension of backward variable selection options provided in the LOGISTIC procedure. The macro is easily implemented in any dataset by having the user specify the IV, DV, additional covariates, and threshold of difference in odds ratio which is used for removal of additional covariates.

INTRODUCTION

Variable selection, or identification of confounders, is a fundamental component of statistical modeling in epidemiology. A number of variable selection procedures have been suggested, such as forward and backward, which are both step-wise methods.^{1,2} Frequently, automated regression procedures employ a step-wise approach based on overall statistical significance for inclusion of covariates using p-value as the metric.³ However, it is commonly agreed that a step-wise approach relying on "change-in-estimate" for covariate inclusion is a superior method for maximizing the relevance of covariates included in the model.^{3,4}

When using logistic regression to model the effect of an exposure of interest (IV) on a binary outcome variable (DV), the "change-in-estimate" procedure examines the percentage change in the adjusted odds ratio (aOR) for the association between the IV and DV upon removal of a particular covariate.¹⁻⁶ The standard convention is a change in the aOR of 10% or more suggests the covariate is important to the model and should be left in, though newer research suggests that a 5% change may be a sufficient cut-off depending on the size of the exposure-outcome relationship.⁵

While a variable selection procedure based on percentage change in odds ratio has many statistical and epidemiologic advantages, implementation is computationally intensive. For instance, if there is a statistical model with a single IV, DV, and 10 additional covariates eligible for removal, as many as 55 separate models are needed to implement this procedure.

We present a macro that automates the covariate selection process using backward variable selection based on "change-in-estimate." The macro enables the evaluation of an arbitrary number of additional covariates with a user-specified threshold for inclusion in the model based on change in aOR upon removal.

MACRO BACKWARD_OR_ELIM

Macro `backward_or_elim` implements the “change-in-estimate” procedure and produces output which thoroughly details every iteration, including all full and reduced models.

Macro `backward_or_elim` procedure

STEP 0: Set macro arguments: IV, DV additional covariates, and threshold for inclusion in the model based on change in aOR upon removal.

STEP 1: Run logistic regression with full set of additional covariates. Compute odds ratio between IV and DV.

STEP 2: Run logistic regression with a reduced set of additional covariates, running a separate model for the full set minus one covariate at a time in a leave-one-out manner. Compute odds ratio between IV and DV in each of these reduced models.

STEP 3: Identify the additional covariate that has the lowest effect on the odds ratio between IV and DV upon removal from the set of additional covariates. If this impact is less than the user-defined threshold, then delete this covariate from the additional covariate set and return to STEP 1. Otherwise, proceed to STEP 4.

STEP 4: End; display final iteration table.

EXAMPLE

The ICU dataset is used to demonstrate macro `backward_or_elim`.^{7,8}

Name	Description	Codes/Values
STA	Vital Status	0 = Lived 1 = Died
INF	Infection Probable at ICU Admission	0 = No 1 = Yes
GENDER	Gender	0 = Male 1 = Female
CAN	Cancer Part of Present Problem	0 = No 1 = Yes
CPR	CPR Prior to ICU Admission	0 = No 1 = Yes

Figure 1: Variables used in ICU dataset

This example is for illustrative purposes only and must not be interpreted to have any scientific relevance. The macro is invoked using:

```
%backward_or_elim(IV=INF, DV=STA, covariates=GENDER CAN CPR, threshold=0.05,  
dataset=Icu_data);
```

Backwards elimination procedure for independent variable INF, dependent variable STA, and additional covariates at threshold 0.05

Iteration	Full Model aOR	Reduced Variable	Reduced Model aOR	Change in aOR
1	2.241	CAN	2.236	0.209%
		CPR	2.502	11.64%
		GENDER	2.242	0.025%
2	2.242	CAN	2.236	0.246%
		CPR	2.506	11.76%
3	2.236	CPR	2.500	11.81%

Output 1: Output from backward_or_elim macro using ICA dataset^{7,8}

ITERATION 1

Iteration 1 corresponds to the model with dependent variable STA, independent variable INF, and three additional covariates (CAN, CPR, GENDER). The full model aOR is the odds ratio between IV and DV adjusted for these three additional covariates, and is equal to 2.241. Note that the IV and DV are fixed for every model in this procedure, and that the DV remains in the model regardless of which additional covariates are included.

In iteration 1 there are three reduced models which are indexed by which additional covariate is removed. The model with removed variable CAN includes the two additional covariates (CPR, GENDER); the model with removed variable CPR includes the two additional covariates (CAN, GENDER); the model with removed variable GENDER includes the two additional covariates (CAN, CPR).

For each of these reduced models, the odds ratio between IV and DV is computed while adjusting for one less additional covariate. The reduced models with the change in aOR less than the threshold of 5% are shown in bold type, and the model with the lowest change in aOR among models eligible for removal is highlighted. The model corresponding to GENDER, which adjusts for (CAN, CPR), has an aOR that is only 0.025% different than the full model aOR, therefore GENDER is removed from the set of additional covariates.

ITERATION 2

Iteration 2 begins with the updated full set of additional covariates (CAN, CPR). Note that the full model aOR in iteration 2 is 2.242, which is the same as the reduced model aOR in iteration 1 for the reduced model which excludes GENDER, since in both models the odds ratio is adjusted for (CAN, CPR).

In iteration 2 there are two reduced models: the model with removed variable CAN including the single additional covariate CPR and the model with removed variable CPR including the single additional covariate CAN. The model corresponding to CAN, which adjusts only for CPR, has an aOR that is only 0.246% different than the full model aOR, therefore CAN is removed from the set of additional covariates.

ITERATION 3

In iteration 3 the set of additional covariates is only CPR. The reduced model aOR corresponds to a model with no additional covariates, and the odds ratio between IV and DV is 11.81% different than the full model aOR. Therefore, the additional covariate CPR is not removed and the procedure ends.

Note that the initial model considered included IV, DV, and the three initial covariates (CAN, CPR, GENDER), while the final model includes IV, DV, and only the additional covariate CPR.

SOURCE CODE

```
%macro backward_or_elim(IV, DV, covariates, threshold, dataset);

/* Initialize variables */
%let cov_list=%sysfunc(compress(&covariates,(,)),);
%let iteration=1;
%let num=%sysfunc(countw(&cov_list));
%let minimum=&threshold;

ods exclude all;
%do %while(&minimum<=&threshold or %eval(iteration<&num));

  /** Step 1: run full model ***/
  ods output OddsRatios=OddsData_Full;
  proc logistic data=&dataset descending;
    class &DV / param=ref ;
    model &DV = &IV &cov_list;
  run;
  ods output close;
  proc sql noprint;
    select OddsRatioEst into :fullor separated by ' '
    from OddsData_Full
    where Effect="%UPCASE(&IV)";
  quit;

  /** Step 2: run reduced models ***/
  %do i = 1 %to %sysfunc(countw(&cov_list));
    %let cov_list_reduced =
%sysfunc(tranwrd(&cov_list,%scan(&cov_list,&i),));
    ods output OddsRatios=OddsData_Reduced;
    proc logistic data=&dataset descending;
      class &DV / param=ref ;
      model &DV = &IV &cov_list_reduced;
    run;
    data OddsData_Reduced;
      length Effect $25 Removed $25;
      set OddsData_Reduced;
      removed="%sysfunc(scan(&cov_list,&i))";
      if Effect ^= "%UPCASE(&IV)" then delete;
    run;
    proc append data=OddsData_Reduced base=OddsData_Merged; run;
  %end;

  /** Step 3: compute effect of deleting one variable on OR ***/
  data OddsData_Merged;
    set OddsData_Merged;
    delta = abs((OddsRatioEst-&fullor)/&fullor);
    iteration=&iteration;
    oddsratio=&fullor;
  run;
  proc sql noprint;
    select min(delta) as minimum, removed into :minimum, :removedvar
    from OddsData_Merged
    having delta=minimum;
  quit;
  data OddsData_Merged;
```

```

    set OddsData_Merged;
    elim = "&removedvar";
run;
proc append data=OddsData_Merged base=OddsData_Final; run;
proc datasets nolist;
    delete OddsData_Merged;
quit;
/** remove &removedvar, the variable with lowest effect on OR */
%if %eval(&minimum<&threshold) %then %do;
    %let cov_list=%sysfunc(tranwrd(&cov_list,&removedvar,));
    %let iteration=%eval(&iteration+1);
%end;

%end;
ods exclude none;

data final;
    set OddsData_Final;
run;

proc datasets nolist;
    delete OddsData_Full OddsData_Reduced OddsData_Merged OddsData_Final;
quit;

ods rtf;
title "Backwards elimination procedure for independent variable &IV,
dependent variable &DV, and additional covariates at threshold &threshold";
PROC REPORT DATA=Final NOWD;
    COLUMNS iteration oddsratio removed oddsratioest delta elim;
    DEFINE iteration / GROUP 'Iteration';
    DEFINE oddsratio / GROUP 'Full Model aOR';
    DEFINE removed / GROUP 'Reduced Variable';
    DEFINE oddsratioest / 'Reduced Model aOR';
    DEFINE delta / FORMAT=Percent8.3 GROUP 'Change in aOR';
    DEFINE elim / GROUP noprint;
    break after iteration/;
    compute after iteration;
    line '';
endcomp;
COMPUTE delta;
    IF (delta<&threshold) THEN DO;
        CALL DEFINE(_col_, "STYLE", "STYLE=[FONT_WEIGHT=BOLD]");
    END;
ENDCOMP;
COMPUTE elim;
    IF (elim = removed and delta<&threshold) THEN DO;
        CALL DEFINE(_row_, "STYLE", "STYLE=[BACKGROUND= cxDDDDDD]");
    END;
ENDCOMP;
RUN;
ods rtf close;

%mend backward_or_elim;

```

CONCLUSION

This flexible macro implements a variable selection technique which is of substantial epidemiologic interest. This current implementation is limited to the case of logistic regression with binary IV, DV, and additional covariates. This macro has already been extended to the cases of:

- categorical IV and DV, and categorical or continuous additional covariates
- adding additional covariates that are not eligible for removal, therefore creating models with an IV, DV, additional covariates eligible for removal, and additional non-removable covariates
- using additional options within PROC LOGISTIC (such as WEIGHT)
- using this framework for any generalized linear regression method in the GLM procedure

These extensions are available by request from the author.

REFERENCES

1. Lee P.H. 2014. "Is a Cutoff of 10% Appropriate for the Change-in-Estimate Criterion of Confounder Identification?" *American Journal of Epidemiology*, 24(2):161–167.
2. McNamee R. 2005. "Regression modelling and other to control confounding." *Occupational and Environmental Medicine*, 62(7): 500–506.
3. Greenland, S. 1989. "Modeling and Variable Selection in Epidemiologic Analysis." *American Journal of Public Health*, 79(3):340–349.
4. Walter S, Tiemeier H. 2009. "Variable Selection: Current Practice in Epidemiological Studies." *European Journal of Epidemiology*, 24(12):733–736.
5. Robins J.M., Mark S.D., and Newey W.K. 1992. "Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders." *Biometrics*, 48(2):479–495.
6. Greenland S, Daniel R, Pearce N. 2016. "Outcome Modeling Strategies in Epidemiology: Traditional Methods and Basic Alternatives." *International Journal of Epidemiology*, 45(2):565–575.
7. Lemeshow, S., Teres, D., Avrunin, J. S., Pastides, H. 1988. "Predicting the Outcome of Intensive Care Unit Patients." *Journal of the American Statistical Association*, 83(402):348–356.
8. Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: John Wiley & Sons.

CONTACT INFORMATION

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

Evan Kwiatkowski
University of North Carolina at Chapel Hill
ekwiatkowski@unc.edu

Hannah Crooke
PAREXEL International and University of North Carolina at Charlotte
hannah.crooke@PAREXEL.com

Kathy Roggenkamp
University of North Carolina at Chapel Hill
kathy_roggenkamp@unc.edu