Robust Standard Error Estimate for Cluster Sampling Data: A SAS/IML® Macro Procedure for Logistic Regression with Huberization

Honghu Liu, Department of Medicine, UCLA, Los Angeles, California

ABSTRACT

Data sets with cluster structure are very common in practical business and research, one has to take into account the intra-cluster correlation in data analysis. This paper systematically discusses the Huber/White standard error estimate for cluster sampling data in logistic regression, and presents a user-friendly SAS/IML® macro procedure which can automatically fit logistic model, calculate robust standard errors and produce confidence intervals for odds ratio. The robust standard error calculated in this procedure also has a finite sample-adjustment feature which is now available in most updated version of some statistical software. The syntax for the procedure is simple and easy. One data example is shown to illustrate how to actually use the procedures. The SAS® system products included in this work are SAS®, SAS/STAT® and SAS/IML® . This procedure can be run on any IBM compatible personal computer, MVS/UNIX system and any other computer platforms with a working SAS® system.

INTRODUCTION

We often encounter data set with cluster structure. For example, subjects collected from different cities or hospitals are nested within cities or hospitals. Observations within city or hospital more likely have similar characteristics. These kind of data have intra-city or intra-hospital correlations embedded in the data structure, which have to be taken into account in parameter estimation.

This work discusses the Huber method, also known as White or Sandwich method, of robust standard error estimate for cluster sampling data in logistic modeling. A user friendly SAS/IML® macro procedure is introduced, which can automatically fit logistic regression model, calculate robust standard error for parameter estimates, test statistical significance of parameter estimates as well as produce confidence intervals for adjusted odds ratio from the fitted model.

A real data example is given to demonstrate in details how to exactly use the procedures in data analysis. The SAS® products to support this macro procedure include SAS®, SAS/STAT® and SAS/IML®. This macro procedure can be run on any IBM compatible personal computer, SUN station, MVS/ UNIX system and any other computer platforms with a working SAS system.

HUBER'S ROBUST STANDARD ERROR ESTIMATE

The traditional standard error estimates for logistic regression models based on maximum likelihood from independent observations is no longer proper for data sets with cluster structure since observations in the same clusters tend to have similar characteristics and are more likely correlated each other. Robust standard error estimates are needed to take into account of the intra-cluster correlation. Huber (1967) has proposed a formula, which is a theoretical (asymptotic) bootstrap or jackknife, for calculating robust standard errors if there is heteroscedasticity, clustered sampling.

Let us denote

$$p_i = \frac{e^{x_i\beta}}{1-e^{x_i\beta}} \qquad \text{i=1,.....n}$$

be the probability of an event, where x_i is the covariates associated with the event probability and β is the regression coefficients.

Let

$$L = \log(\prod f(x_i)) = \sum \log(f(x_i)) = \sum l(x_i)$$

i=1,.....,n. be the log likelihood, then define score function to be

$$S_i = \frac{\partial(L)}{\partial(x_i\beta)}$$

and the Hessian be

$$H_j = \frac{\partial^2 L}{\partial(x_i\beta)^2}$$

for the i[th] observation, i=1,.....,n.

Suppose that we drop the i[th] observation from the model, then the estimates would shift by the amount of $-D^{-1}S_i x_i^T$ where the matrix $D = \sum_i H_i(x_i^T x_i)$. We assume that no single observation has very large effect in the fitting, then the effect of dropping two observations is roughly the sum of the effect of dropping each observation individually. The same token is true for observations from a cluster, dropping all the members of a group is roughly equivalent to the sum of dropping each member in turn. By Huber's formula, the robust standard variance estimate is:

$$Var(\beta) = D^{-1}(\sum_i S_i x_i^T x_i S_i)D^{-1}. \qquad (1)$$

we can see from the physical appearance of the above formula that why people also name the estimate as "Sandwich Estimates".

For logistic model, we can, after some algebra, show that the score is

$$S_i = \frac{\partial(L)}{\partial(x_i\beta)} = y_i - p_i$$

and the Hessian is

$$H_j = \frac{\partial^2 L}{\partial(x_i\beta)^2} = p_i(1-p_i).$$

Then plug into (1), we have

$$Var(\beta) = (\sum_i (p_i(1-p_i)(x_i^T x_i)))^{-1} *(\sum_i (y_i - p_i)x_i^T x_i (y_i - p_i)) *(\sum_i (p_i(1-p_i)(x_i^T x_i)))^{-1}.$$

Suppose that we have C clusters (each cluster has g_j observations, j=1,...,C), and each cluster is independent with other clusters. Then the robust standard error is

$$Var(\beta) = \tilde{D}^{-1}(\sum_j U_j^T U_j)\tilde{D}^{-1}. \qquad (2)$$

where $\tilde{D} = \sum_{j=1}^{C}\sum_{k=1}^{g_j}(H_{jk}x_{jk}^T x_{jk})$

$U_j = \sum_{k=1}^{g_j} x_{jk}S_{jk}$, j=1,...,C., the contribution to score from each cluster

For data with weights, we will have $X^T X$, $X^T y$ and $y^T y$ replaced by $X^T WX$, $X^T Wy$ and $y^T Wy$, where W is a diagonal matrix whose diagonal elements are the elements of w, the vector of weights.

Since the robust standard error estimate is based on sample data, a finite sample correction is necessary to adjust the estimates more closer to the population value. There are two popular adjustment, one is the regression-like formula (Fuller et al. 1986),

$$q_c = \frac{N-1}{(N-p-1)} * \frac{C}{C-1}$$

where N is the number of total observations and p is the number of predictors in the model (without counting the constant 1 for intercept) and C is the number of clusters in the data. The second adjustment is the asymptotic-like formula:

$$q_c = \frac{C}{C-1}.$$

For data set with very large number of observations, the effect of these two adjustment are about the same. Taking into account the finite sample correction, the formula for the Huber robust variance is:

$$Var(\beta) = q_c\tilde{D}^{-1}(\sum_j U_j^T U_j)\tilde{D}^{-1}.$$

A SAS/IML® MACRO PROCEDURE

This SAS® macro procedure for calculating Huber robust standard error in logistic model is written by SAS®, SAS/STAT® and SAS/IML®. The procedure calls the SAS® PROC LOGIT in the program and uses it to produce coefficient estimates for logistic regression (these coefficient estimates are still consistent even with cluster data structure). Intermediate values needed to conduct the calculation of robust standard errors such as the total number of observations, the number of clusters, the number of predictors, a character string containing all the

independent and dependent variable names with comma as delimitations, and some data manipulations such as to get rid of all the observations with missing values on either dependent or independent variables are calculated or conducted by using macro language in SAS®. The core part of the calculation for robust standard errors is carried out by using SAS/IML®.

To have a closer look into the macro procedure, let us take a piece. As an example, let us look at the part which creates a macro variable named *NP* containing the number of predictors and the macro variable named *XYM* containing all the predictors and the dependent variable with comma as delimitation, the code is

```
%macro npm;
%global np xym;
%let np=1;
%let xm=;
%let xs=;
%let xn=&x;
%do %until(&xs=&xn);
%let xs=%scan(&xn,1,' ');
%if &np=1 %then %let xm=&xs;
%else
%let   xm=&xm%str(,)&xs;
%let np=%eval(&np+1);
%let l=%eval(%length(&xs)+2);
%if &xs=&xn %then %goto done;
%let xn=%substr(&xn,&l);
%end;
%done:
%let xym=&y%str(,)&xm;
%mend npm;
```

At beginning of the procedure, the working data set containing the dependent variable, all the independent variables, cluster id and weights, if any, is identified by the data set name parameter entered through the macro procedure syntax or by the system variable _last_ if no data set name entered (using the default value, i.e., the last active data set). Then, the procedure automatically calculate those intermediate parameters needed in later calculation, such as the total number of observations, the number of clusters and number of predictors (see the above macro code). These intermediate parameters are not asked from user, but rather computed directly from the given data in the effort trying to make the procedure syntax more user-friendly. The number of predictors which is used later to determine the matrices dimension in the core calculation and the character string of all independent variables and dependent variable names are obtained through macro variable operation with quoting features (see the above macro code). The macro code for this part can work on any length of variable listing with spaces in between. The code scans over the variable string each time, and find the first blank, then gets the sub-string of a variable name (from the first character up to right before the first blank). This part of code itself is an internal sub-macro procedure and it operates very efficiently.

A data set with no missing values on the dependent variable and all the independent variables is created for all the calculation activities in the procedure. Because SAS PROC LOGIST only uses the observations without any missing values on both dependent and all independent variables in its computation, a data set without any missing values has to be created to ensure that all the calculations for the robust standard errors and the coefficients parameter estimates are both based on the same observations.

The calculation of the robust standard error is based on the Huber's formula (see section II). The Score and Hessian values for each observation are created right after the invoking of PROC LOGIST.
The finite-sample adjustment in the procedure is based on Fuller's method which takes into account the number of observations, the number of clusters and the number of independent variables. If weight is used in coefficient parameter estimation, then probability weights is used in robust standard error calculation. There are five levels of choices for confidence intervals of odds ratio, they are 80%, 85%, 90%, 95% and 99%. These confidence intervals are calculated based on asymptotically normal theory.

The outputs of this procedure include logistic model fitting information, parameter estimations, robust standard errors, testing statistics and p-value of the parameter testing results based on robust standard errors, and also confidence intervals of odds ratio.

There is a program heading at the beginning of the procedure code in which a description of the

procedure and the required syntax has been described. The required syntax for the procedure is:

%hlogist(*x_list,y,cluster_id,wt,ci,dataset*);

where *x_list* is the list of all the independent variables with space in between.
y is the dependent variable.
cluster_id is the id for cluster.
wt is the weight, for un-weighted analysis, enter value 1.
ci is the level of confidence interval for odds ratio, the choices available are 80,85,90,95 and 99.
dataset is the name of the data on which you will run the analysis. To be consistent with SAS® procedure convention, the default is the last active data set (e.g., if leave 'data set' as blank in one's input).

To use the procedure, one first needs to include the entire code of the procedure to his program by copying it into your program or by using %include statement to load the procedure in, then issue the macro statement:

%hlogist(*x_list,y,cluster_id,wt,ci,dataset*);

with all the parameters properly substituted.

DATA EXAMPLE

In this section, we apply the SAS/IML® macro procedure to a data set about end of life Cardiopulmonary Resuscitation (CPR) preference for 3137 old hospitalized patients. The dependent variable indicates that patient wanted CPR or DNR(do not resuscitate) at month two after the study entry. It is of interesting to health care workers to know what patient characteristics predicting CPR preference. Since these 3137 patients were from five different medical centers of different locations, there is a possible intra-location correlation among the patients, and we need to take this into account in our analysis. Independent variables chosen for modeling CPR preference are patient gender (baseline female), age, education level, marriage status, whether the patient lives alone, race ethnicity (baseline white), quality of life measured at the study entry, depression score, functional status at the study entry and two month survival prediction. The table 1 below is the logistic model fitting results on CPR preference without adjusting medical center clustering. The first column is the variable names, the second column is the coefficient parameter estimates, the third column is the standard error of the point estimate and the last column is the p-value for testing the point estimate.

TABLE 1
Logistic model for CPR preference
(without adjusting cluster)

Variable	Estimate	S. E.	P_value
Intercept	4.087	.6512	.0001
Male	0.541	.1325	.0001
Age	-0.010	.0046	.0292
Education	0.028	.0217	.2006
Marriage	-0.061	.1595	.7032
Live alone	-0.466	.1751	.0079
Black	0.940	.2481	.0002
Other race	0.753	.3780	.0463
Quality of life	0.004	.0742	.9618
Depression	-0.212	.0935	.0235
Functional status	0.191	.0508	.0002
Survival prob.	-2.572	.4115	.0001

Now let us take into account the intra-location correlation by using the SAS/IML® macro procedure. Issue at beginning of the program

%include '*hlogist.mac*';

to load in the macro procedure ('*hlogist.mac*' is the file name of the SAS/IML® macro procedure). Then issue the following macro statement to invoke and run the procedure (the working data set we used is named *cprdata*, the dependent variable is *cpr*, the cluster variable is *m_center* and we want 95% confidence interval for the odds ratio):

%hlogist(*male age edu marriage alone black otherace q_life depress funct surv, cpr, m_center*,1, 95, *cprdata*);

Notice that we entered '1' for parameter of weight since we are doing unweighted analysis (equivalent as weight equal to 1). After the procedure has run, we will have all model fitting information from original SAS® PROC LOGIST and plus robust standard errors estimates, testing statistics and p_value based on the robust standard errors and confidence intervals

of odds ratio. The following Table 2 is the summarized results

TABLE 2
Logistic model for CPR preference
(After adjusting cluster)

Variable	Estimate	R. S. E.	P_value
Intercept	4.087	.7089	.000
Male	0.541	.0818	.000
Age	-0.010	.0053	.058
Education	0.028	.0269	.302
Marriage	-0.061	.1971	.758
Live alone	-0.466	.0969	.000
Black	0.940	.2995	.002
Other race	0.753	.4169	.071
Quality of life	0.004	.0570	.950
Depression	-0.212	.0852	.013
Functional status	0.191	.0928	.040
Survival prob.	-2.572	.4488	.000

Compare with Table 1, we can see that the robust standard errors (RSE) are larger than the regular standard errors for most of the variables except for male, alone, quality of life and depression. In this case, most of the confidence intervals for odds ratio became wider.

DISCUSSION

Huber robust standard error estimates are very important and useful for statistical analysis in business and research since people very often encounter data sets with cluster structure and have to take intra-cluster correlation into account in their statistical analysis. Without robust standard error, one could end up with wrong conclusion about statistical testing on point estimates.

SAS® is one of the most powerful statistical software, it deserves to have the ability to estimate robust standard errors for logistic modeling. This procedure is easy to use and it requires only few parameters to be entered. The outputs give users not only the robust standard errors, but also the whole logistic model fitting information, the statistical tests of the parameter estimates based on the robust standard errors as well as the confidence intervals of odds ratio. The limitation of the procedure is that it is designed only for binary outcome, it does not work for ordinal outcomes with more than two levels.

REFERENCE

Agresti, A. (1984), Analysis of Ordinal Categorical Data, New York: John Wiley & Sons, Inc.

Fuller, W. A., W. Kennedy, D. Schnell, G. Sullivan, H. J. Park. 1986. PC Carp. Ames, IA: Statistical Laboratory, Iowa State University.

Huber, P. J. 1967. The behavior of maximum likelihood estimates under non-standard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1: 221-233.

SAS/IML® Software: Usage and Reference, Version 6, SAS® Institute Inc., Cary, North Carolina.

SAS/STAT® User's Guide, SAS® Institute Inc., SAS® Campus Drive, Cary, North Carolina.

SAS® Guide to Macro Processing, Version 6, Second Edition, SAS® Institute Inc., Cary, North Carolina.

STATA® Statistical Software, Stata Corp. 1997. Release 5.0 College Station, TX: Stata Corporation.

White, H. 1980. A heteroskedasticity-consistent covariace matrix estimator and a direct test for heteroscedasticity. *Econometrica* 48: 817-830.

Honghu Liu
General Internal Medicine & Health Service Research
UCLA Department of Medicine
LA, CA 90095-1736
Phone: 310-794-7396
Fax: 310-206-0719
email: hhliu@ucla.edu