Paper 81-2018

An Efficient Way to Output Univariate Analysis Results with Multiple Predictors

Bocheng Jing, Northern California Institute for Research and Education, California

Kathy Z Fung, San Francisco VA Medical Center, California

W John Boscardin, University of California, San Francisco, California

Abstract

When building a predictive model, researchers often begin the selection process with a univariate regression analysis for each variable. The variable with a significant univariate test result that is under a pre-specified cut-off level will potentially be the candidate in the final model. If a study contains many variables, it is repetitive to run the same univariate regression procedure for each predictor and outputting the analysis results after each iteration would be tedious and cumbersome. Using DO LOOP can fulfill the task of procedure automation; however, it is time costly and inefficient. In this paper, we will use a method that runs the univariate regression in an efficient way. We will also use SAS® Output Delivery System (ODS) to deliver the analysis results of interest (e.g. p-value, odds ratio, C-statistics) in integration.

Introduction

Univariate tests are crucial for model setup. Not only because they can help researchers to determine the potential predictors for the final model, but they can also help examining the best definition of each variable, such as if a variable should be dichotomous, continuous, discrete, polychotomous, etc. When running a univariate model for each variable, the most common method is to use a DO LOOP, in which we pre-specify variable names, then loop them into the model. If you have 100 comorbidity variables that are needed to be examined, the time spent to run the overall univariate analysis would be 100 times longer with each model iteration. Moreover, the results of each iteration will be displayed separately, which is difficult for researcher to review the output systematically.

```
/*Conventional DO-Loop*/
 %let X1=variable_1;
 %let X2=variable_2;
 ...
 %let Xn=variable_n;

 %macro Uni_logistic(Mydata, Varcounts);
%do i = 1 %to &Varcounts;
   proc logistic data=&Mydata;
```

```
   class Y(ref='0')
            X[i](ref='0')/param=ref;
   model Y = X[i];
%end;
%mend;
```

By-Group Processing

A quick way to build univariate regression analysis is to use BY-group processing. Instead of running the same process multiple times, the BY statement runs the regression within each BY group, which integrates the running process. It outputs the results as an integration by the ODS.

1. Transform the data

The final dataset usually culminates with a subject identifier, the outcome variable, and many predictors. We often refer this format as "wide format." To use the BY-group processing, we need to transform the data from -wide format to long format. In table 1 below, the wide format contains patients' IDs, outcome Y, predictors X1, X2, and X3 (all predictors are binary).

Table 1. Wide Format Final Dataset

ID	Y	X1	X2	X3
1	1	1	0	1
2	0	1	1	0
3	1	0	0	1

Table 2 indicates the long format of Table 1, which only contains ID, output, predictor (predictor values), and a predictor **Table 2. Long Format Final Dataset**

ID	Y	X (predictors values)	X_indicator
1	1	1	X1
2	0	1	X1
3	1	0	X1
1	1	0	X2
2	0	1	X2
3	1	0	X2
1	1	1	X3
2	0	0	X3
3	1	1	X3

The following step transforms all the predictor variables from wide format to long format. The outcome variable is dropped and will be merged back after the predictor variables transformation.

```
/*Transform data from Wide format to Long format*/

proc transpose data=wide_format (drop=Y)
```

```
   out=long1(rename=(Col1=Value)) name=X_ind;
   by ID;
   var X:/*Speficy all the variables name.
            : will be used when the variables have the same
prefix*/;
run;


/*Merge back with the outcome and SORT DATA by predictor
indicator*/
proc sql;
     create table long_format as
     select A.ID, A.Y, B.*
     from wide_format A inner join long1 B on A.ID=B.ID
     order by B.X_ind;
quit;
```

One essential key to perform BY-group analysis is to sort the data by the BY-group variable. The long format dataset has been sorted by the predictor indicator.

2. Univariate Regression Analysis

With the long format dataset, we can start the BY-group processing. Since the outcome is binary, logistic regression will be used for the univariate analysis.

```
proc logistic data=long_format;
     class Y(ref='0')
           Value(ref='0') /param=ref;
     by X_ind; /*BY-group processing*/
     model Y=Value;
run;
```

3. Real Example

In my current study, 125 disease groups need to be examined to explore the relationship of the outcome: death/function vs. alive/well. P-value, c-statistics and odds ratio needs to be reviewed for each disease group before the final model. Baseline variables age and gender are also adjusted for each regression. Both do loop and BY-group processing have been used to execute the 125 of logistic regression models:

Outcome (death/function vs alive/well) ~ age + gender + disease group[i].

The do loop method took approximately 21.89 seconds while the BY-group processing took approximately 11.97 seconds. The difference will increase significantly as the sample size and the number of variables increase.

Output the results via ODS

SAS® procedure often generates much more statistics than one wishes to review. Running the univariate regression analysis will generate the same unwanted statistics through each BY-group. All the univariate results of each BY-group will be displayed in the Result window, making it hard for researchers to review systematically. The SAS® ODS system provides a unique way to capture the statistics of interest. In this paper, our statistics of interest would be the variable's point of estimate, p-value, odds ratio with its confidence interval and c-statistics for each univariate regression. By using ODS, we can generate a summary table for the statistics above, which makes it much easier to review.

```
/*Using ODS to output summary statictics*/
ODS trace on;
proc logistic data=long_format;
     class Y(ref='0')
            Value(ref='0') /param=ref;
     by X_ind;
     model Y=Value ;
run;
ODS TRACE OFF;

ODS output parameterEstimates=pest /*Output point estimation and
p-value*/
```

```
              Oddsratios=OddsR /*Output Odds ratio and CI*/
              Association=ASSO; /*Output C statistics*/
proc logistic data=long_format;
     class Y(ref='0')
            Value(ref='0') /param=ref;
     by X_ind;
     model Y=Value ;
run;
ODS output close;

/*Merge all output data as a summary table*/

data pest1 (keep=X_ind estimate probT rename=probT=p_value);
     set pest (where= (Variable="Value"));
run;

data OddsR1 (keep=X_ind OddsratioEst LowerCL UpperCL);
     set OddsR;
run;

data ASSO1(keep=X_ind cValue2 rename=cValue2=C_statistics);
     set ASSO(where=(Label2='c'));
run;

data univ_summary_stats;
     merge pest1 OddsR1 ASSO1;
     by X_ind;
run;
```

The ODS TRACE statement displays the analysis results in the SAS® Log, and Output ODS statement creates tables to save all the statistics of interest. When all the statistics are output into tables, we can merge them into a summary table.

Table 3: Summary statistics for Univariate Regression Analysis

	X_ind	Estimate	OddsRatioEst	LowerCL	UpperCL	C_statistics
1	X1	-0.8109	0.444	0.035	5.581	0.600
2	X2	-1.7918	0.167	0.010	2.821	0.700
3	X3	-1.7918	0.167	0.010	2.821	0.700

Conclusion

Using BY-group processing can save the running time of each iteration when building univariate regression analysis for predictive modeling. This method is especially useful when the count of variables and sample size are large. SAS® ODS provides a convenient way to obtain the statistics of interest which helps researchers reviewing potential variables systematically before making decisions for the final model. Both tools are efficient and convenient when dealing with modeling building and variable examination in research field.

Reference

1. Long, S., Abolafia, J., Park, L. “Using SAS® ODS to extract and merge statistics from multiple SAS

procedures into a single summary report, a detailed methodology.” Paper 261-31. Durham, NC. Available at http://www2.sas.com/proceedings/sugi31/261-31.pdf.

2. Wicklin, Rick. “Simulation in SAS: The slow way or the BY way”. Accessed July 18, 2012. Available at: https://blogs.sas.com/content/iml/2012/07/18/simulation-in-sas-the-slow-way-or-the-by-way.html .

Contact Information

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

Bocheng Jing

Northern California Institute for Research and Education

415-221-4810 ext. 24179

Bocheng.jing@ncire.org

www.ncire.org

Kathy Fung

San Francisco VA

415-221-4810 ext.24953

Kathy.fung@va.gov

John Boscardin

University of California, San Francisco

John.boscardin@ucsf.edu