

Paper SA05_05

Using ODS to Perform Simulations on Statistics from SAS® Procedures

Andrew A. Kramer, Ph.D., Cerner Corporation, Vienna, VA

ABSTRACT

The Output Delivery System (ODS) is SAS' mechanism for exporting output from procedures to a variety of format destinations. One of the format destinations is a SAS data set. Thus the statistics of interest from a SAS procedure such as PROC LOGISTIC can be captured. If the SAS procedure is repeated many times then a SAS data set containing the results from that procedure can be iteratively built. With the use of a few macro variables and a single ODS statement simulation studies of output statistics can be developed. An example is given here using PROC LOGISTIC to perform a simulation study of the behavior of the Hosmer-Lemeshow GOF statistic (HL-GOF) when large data sets are used. A logistic regression of 50,000 observations was run 1,000 times with the HL-GOF statistic being saved in a data set. A test of the distribution of the 1,000 HL-GOF statistics confirms that this statistic is distributed as a Chi square variable with 8 degrees of freedom.

INTRODUCTION

Starting in version 7 SAS introduced the Output Delivery System (ODS). The major benefit of this new feature was the ability to direct output to a number of different format destinations. Output could be sent to an HTML file, Word document, PDF file, etc... However ODS also permitted the rerouting of output to a SAS dataset. Thus statistics that previously had to be extracted by hand from a SAS listing could now be sent directly to a file. Examples include the coefficient of determination from a linear regression, F-test value from an Analysis of Variance table, p-values for tests of hypotheses as well as many other statistics from SAS procedures.

ODS allows exporting of statistics by way of table elements. Each SAS procedure has one or more ODS tables associated with it. The tables have an ODS Table Name and contain statistics from a statement or option in that procedure which can be exported. PROC LOGISTIC has 46 tables that contain elements that can be exported to a SAS data set. This is accomplished by invoking a table via an ODS statement. As an illustration first consider the SAS statements below:

```
proc logistic data = hospital outest = parm_values;  
    SESUG:    model mortality (event='1') = age ventilator wbc sbp /lackfit;  
    output out = hosp_results p=pred;  
    title "Logistic Regression of Age, Ventilated (Y/N), White Blood Cell count,  
          and Systolic Blood Pressure on Mortality";  
run;
```

The above statements output the estimated coefficient values from a logistic regression procedure to a SAS dataset called parm_values. The Hosmer-Lemeshow Goodness-of-Fit (HL-GOF) statistics are displayed in the SAS listing window, and the original dataset along with the value predicted by the logistic regression model for each observation are output to a SAS dataset called hosp_results. But suppose that instead of sending the HL-GOF statistics to the SAS listing window we wanted to export these statistics to a SAS data set. All we need to do is to add an ODS statement to the logistic procedure.

```
ODS output LackFitChiSq = HLchi ;
```

The ODS statement above retrieves the SAS-named ODS Table LackFitChiSq, which has elements for the overall Chi Square value from the Hosmer-Lemeshow test and the associated p-value. The values for these elements are sent to a SAS data set called HLchi (or any other name that we wish).

INCORPORATING ODS OUTPUT FROM PROC LOGISTIC INTO A SIMULATION

The HL-GOF statistic is ostensibly distributed as a Chi square variable with 8 degrees of freedom (d.f.) under the null hypothesis (Hosmer & Lemeshow, 1980; Lemeshow and Hosmer 1982). This is not intuitive as we expect that a test using deciles would have 9 d.f., that is, the number of groups (10) minus 1. Hosmer & Lemeshow maintain that simulation studies indicated that 8 d.f. was correct. Their simulations were run on small data sets (≤ 200 observations) with a logit containing only a few variables. It would be reassuring to see if the same distributional properties are maintained when studying much larger sample sizes with a complex logit that had multiple integer and

continuous variables (Kuss, 2002; Pulkstenis & Robinson, 2002). One possible approach is to run 1,000 simulations of a complex model that was estimated using 50,000 observations. The ODS feature in SAS allows one to route statistics into a file and thus carry out such a simulation study.

The first step is to create a SAS data set with a single “starter” observation for the H-L results and turn off all printed log and output lines as they can get very large.

```
data HL; input @1 simnum 4.0 @6 ChiSq 5.3 @12 ProbChiSq 6.4;
datalines;
0000 0.000 0.0000
;
run;

options nonotes nosource 2 errors=0; * Turns off printing of log file & errors;

ods listing close; * Turns off printing of output;
```

Next in order to run the simulation study we create a data set containing 50,000 observations with the values of the independent variables. For this example I chose a complex logit function that consisted of 20 binary variables and 3 continuous variables. This data set will have all variables set = 0 and will serve as the “setup” data set for the simulations.

```
data setup; input x1-x20 z1-z3;
datalines;
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.00 0.00 0.00
;
do i = 1 to 50,000; output; end;
drop i;
run;
```

Now we’re ready to carry out the 1,000 runs of the simulation. This requires a couple of macros as shown below. The first macro takes the “setup” data set and replaces the zero values with randomly acquired values based on a specific type of equation. For this study I chose a complex logit function that consisted of the 20 binary variables and three continuous variables as independent variables. I could have chosen any particular structure for the logit but I wanted to use here a logit that is similar to one found in medical studies predicting hospital mortality for critically ill patients. The binary variables were set = 1 if they were greater than $0.03 \cdot i$, with $i = 1$ to 20. Thus there are 20 binary variables with probabilities of 3%, 6%, 9%,60%. The three continuous variables were drawn from a normal distribution with mean = 0.5 and variance equal to 0.1, 0.5, and 1.0, respectively. I set alpha (the intercept) = -4.60 so that the estimated probability of an outcome = 1 was 14%. This all takes place within a macro called **SIMOBS**. The variable simnum is the particular simulation number (1,2,3.....1,000).

```
%MACRO SIMOBS(simnum);
data predict&simnum; set setup;
array binvars(20) X1-X20; * 20 binary variables;
array contvars(3) Z1-Z3; * 3 continuous variables;

do numtrials = 1 to 50,000;
do x = 1 to 20;
if ranuni(-1) le x*0.03 then binvars(x) = 1; else binvars(x) = 0;
end;
Z1 = 0.5+rannor(-1)*0.1; Z2 = 0.5+rannor(-1)*0.5; Z3 = 0.5+rannor(-1);

* creating the logit;
logit = -4.60 + SUM(OF X1 X3 X5 X7 X9 X11 X13 X15 X17 X19)
- 0.5*SUM(OF X2 X4 X6 X8 X10 X12 X14 X16 X18 X20)
+ (Z1-Z2+Z3);

pred = exp(logit)/(1+exp(logit));
```

```

* creating the outcome based on the logit;
observed = 0; if ranuni(-1) < pred then observed = 1;

output;
end;

drop numtrials x;
run;
%MEND;

```

The next step is to run a logistic regression on the data set which was just created and use an ODS statement to output the results to a SAS data set.

```

%MACRO LOGIST(simnum);

proc logistic data = predict&simnum;
  SESUG2005: model observed (event='1') = X1-X20 Z1 Z2 Z3 / lackfit;
  title "Logistic Regression of 50,000 Simulated Observations";
  ODS OUTPUT LackFitChiSq = HL&simnum ;
run;

data HL; set HL HL&simnum; run;

proc datasets; delete predict&simnum HL&simnum; run;

%MEND;

```

Then we write a few lines of macro code to drive the simulations:

```

%MACRO SIMRUN;
  %do i = 1 %to 1000;
    %simobs(&i)
    %logist(&i)
  %end;
%MEND;

```

After **SIMRUN** has finished we'll get rid of the dummy simulation run and turn printing back on.

```

Data HL; set HL; if simnum ne 0; run;

options notes source errors=5;      * Turns back on printing to the log file;

ods listing;                        * Turns back on printing of output;

```

RESULTS OF THE SIMULATION ON HOSMER-LEMESHOW GOODNESS OF FIT

The above simulation code was run. This produced the SAS data named HL that consisted of 1,000 observations with the variables simnum (simulation number), ChiSq, and ProbChiSq. In order to see the distributional properties of the HL-GOF statistic PROC UNIVARIATE was run using the following code.

```

proc univariate data = HL;
  var ChiSq;
  probplot ChiSq / gamma(alpha=est theta=0 sigma=est); *estimate moments of;
  inset mean median stdmean gamma;                  * a Gamma distribution;
  title "Univariate Statistics of HL Stats";
run;

```

The results are given in the table below. The mean HL-GOF was 7.89 and the standard deviation was 3.91, close to the expected mean value of 8.00 and standard deviation of 4.00, respectively, under a χ^2 distribution with 8 degrees of freedom. The probability for lack of fit of a Gamma(4,2) distribution (i.e. a χ^2 distribution with 8 df) was not significant. Finally when the Gamma distribution parameters were estimated by PROC UNIVARIATE using its probplot statement the parameters of 4.04 and 1.95 were very close to the expected values of 4 and 2.

Number of Trials = 1,000		
Number of Observations per Trial = 50,000		
	Expected Value Under χ^2_{8df}	Observed Value
Mean	8.00	7.89
STDEV	4.00	3.91
Gamma(4,2) p-value	>0.05	>0.250
Alpha	4.00	4.04
Sigma	2.00	1.95

The results above strongly suggest that the HL-GOF test does indeed follow a Chi square distribution with 8 d.f. even when the sample size is very large.

CONCLUSION

The ODS in SAS allows the user to capture statistics of interest from any PROC. Utilizing ODS to extract statistics from PROC LOGISTIC made a simulation of the Hosmer-Lemeshow statistic feasible. The technique presented here can be extended to study virtually any statistic generated by a SAS procedure.

REFERENCES

Hosmer, DW and Lemeshow, S (1980), "Goodness of Fit Tests for the Multiple Logistic Regression Model", *Communications in Statistics – Theory and Methodology*, A9(10), 1043-1069.

Lemeshow, S and Hosmer, DW (1982), "A Review of Goodness of Fits Statistics for Use in the Development of Logistic Regression Models", *American Journal of Epidemiology*, 115, 92-106.

Kuss, O (2002), "Global Goodness-Of-Fit Tests in Logistic Regression with Sparse Data", *Statistics in Medicine*, 21, 3789-3801.

Pulkstenis, E and Robinson, TJ (2002), "Two Goodness-Of-Fit Tests for Logistic Regression Models with Continuous Covariates", *Statistics in Medicine*, 21, 79-93.

CONTACT INFORMATION

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

Andrew A. Kramer, Ph.D.
Cerner Corporation
1953 Gallows Road, Suite 570
Vienna, VA 20218
Work Phone: 703-245-8147
Fax: 816-936-7447
Email: akramer@cerner.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.