Making Data Useful Again: The Full Monte

Suzanne L. Cross, The Pennsylvania State University, University Park PA
Tim Leso, The Pennsylvania State University, University Park PA

ABSTRACT

Scads of data are hidden in databases and ignored except for an occasional query. With our application we will bring data back to life and show how rows from several transactional tables can be transformed into useful objects. This paper describes the methods we use to create customized reports using generic macros for object-based statistical analysis, interpretation and presentation of the results using SAS/GRAPH® and ODS LAYOUT.

INTRODUCTION

The data warehouse is a rich source of data for research. Unfortunately, the format of the warehouse puts it out-of-reach for faculty with only spreadsheet skills. It is not easy to transform the rows and columns of the numerous tables into a nice dataset that behaves the way we expect for longitudinal research. We have put a lot of effort into creating extracts, forcing rows into arrays and editing in the data step because it has been our longstanding belief that data, to be useful, must be stored on our PCs where it can be sorted, merged, analyzed and summarized in a unique report. As the warehouse has grown in popularity and research questions proliferated, our PCs have become cluttered with datasets, programs, macros, tables and graphs, and we have come to realize that the answers to these questions are not unique. In practice, most of our analyses are pretty similar—including the interpretation and reporting. Our solution is to modularize our programs and automate the reporting process.

Using the SQL procedure, our favorite SAS/STAT® and SAS/GRAPH procedures, and ODS LAYOUT from SAS® version 9.1, we are designing one-size fits all reports for several scenarios beginning with the simplest, a two by two table.

ONE SCENARIO

An innovative faculty member uses new techniques to teach students in a course and asks the question, "What are the odds that students passing this course will be successful in a subsequent course?" This is a common question and can be automated to the point of prompting for the name of two courses. Our approach is to break-up the program modules into small reusable pieces that accomplish several tasks:

- Transformation
- Data Analysis
- Graphics
- Interpretation
- Presentation

Our goal is to create a "push-button" one-page report that does not clutter our PC with used datasets, programs, gifs and other stuff.

TRANSFORMATION

The original table in the warehouse contains a row for every course registration. We are interested in creating one row for each student containing a grade and a semester for two courses using the last instance of the prerequisite course and first instance of the subsequent course.

.

ID	COURSE	GRADE	SEMESTER
1	SAS 101	D	FALL
1	SAS 101	B	SPRING
1	SAS 102	C	SUMMER
2	SAS 101	B	FALL
2	SAS 102	A	SPRING

ID	COURSE1 GRADE	COURSE2 GRADE	SEMESTER1	SEMESTER2
1	B	C	SPRING	SUMMER
2	B	A	FALL	SPRING

Only three macro variables are used to select two courses and a beginning semester to produce a one-page report.

```
%let course1=%str('SAS  101 ');
%let course2=%str('SAS  102 ');
%let semester=%str('200001FA');
```

To transform the original table in the warehouse, we find all pairs of the selected courses using the SQL procedure to join the table with itself. The result is a temporary SAS dataset having one row for each prerequisite /subsequent course pair.

```
PROC SQL;
connect to odbc(
"DSN=Student;" ||
"Description=NESUG Tables;" ||
"UID=xxx;" ||
"PWD=xxx;"||
"WSID=xxx;" ||
"Address=WAREHOUSE.xxx.xxx.xxx");

create table allPairs as
 select numb_stud_id,
        prereq_code_enrl_grade,
        prereq_code_univ_yr_sem,
        prereq_code_crse_key,
        subseq_code_enrl_grade,
        subseq_code_univ_yr_sem,
        subseq_code_crse_key
  from connection to odbc
 (select
         prereq.code_enrl_grade as prereq_code_enrl_grade,
         subseq.code_enrl_grade as  subseq_code_enrl_grade,
         prereq.code_univ_yr_sem as prereq_code_univ_yr_sem,
         subseq.code_univ_yr_sem as subseq_code_univ_yr_sem,
         prereq.code_crse_key as prereq_code_crse_key,
         subseq.code_crse_key as subseq_code_crse_key,
         prereq.numb_stud_id as numb_stud_id
          from dbo.transcript prereq,
               dbo.transcript subseq
       where (prereq.code_crse_key = &course1
        and   subseq.code_crse_key = &course2
             and   prereq.code_univ_yr_sem = &semester
             and  prereq.numb_stud_id=subseq.numb_stud_id)
       );
disconnect from ODBC;
quit;
```

The SQL procedure doesn't have a method comparable to the automatic first. and last. variables in a sorted data step. Below, the temporary dataset containing all pairs is sorted to keep only one record with last prerequisite course and first subsequent course in a data set called pairs. The data step which previously monopolized our programs is used only to select the last prerequisite course and first subsequent course.

```
proc sort data=allPairs;
    by numb_stud_id prereq_code_univ_yr_sem subseq_code_univ_yr_sem prereq_code_crse_key
subseq_code_crse_key ;

data Pairs ; set allPairs;
    by numb_stud_id prereq_code_univ_yr_sem subseq_code_univ_yr_sem prereq_code_crse_key
subseq_code_crse_key ;
if last.prereq_code_univ_yr_sem and first.subseq_code_univ_yr_sem;

label prereq_code_enrl_grade="Prerequisite Course"
      subseq_code_enrl_grade="Subsequent Course";
run;
```

DATA ANALYSIS

In our example, the frequency procedure in the macro *twobytwo* is used to calculate the row percents, odds ratio and exact confidence limits. These are output into two temporary datasets, freqs and stats, and later processed by the Tabulate procedure for presentation. The two macros, *table1* and *table2,* are called in the final presentation where ODS Layout statements place Table 1 and Table 2 onto the one-page report.

```
%macro twobytwo;
ods listing close;
ods select CrossTabFreqs RelativeRisks;
proc freq  noprint;
   format prereq_code_enrl_grade  subseq_code_enrl_grade  $grd.;
   tables prereq_code_enrl_grade*subseq_code_enrl_grade/
    nopercent  nocol nocum outpct out=freqs chisq relrisk;
   exact pchi or;
   output out=stats or ;
run;
%mend twobytwo;
%twobytwo

%macro table1;
ods proclabel="Two by Two Table";
ods printer style=styles.barrettsblue;
proc tabulate data=freqs   ;
   table prereq_code_enrl_grade=&&course1*
        ('Number  ' pctn<subseq_code_enrl_grade All>='% of row'*
           [style=[font_weight=bold  cellwidth=1in]] ),
           subseq_code_enrl_grade=&&course2 All="Total";
    freq count;
    class prereq_code_enrl_grade/style=[cellwidth=1in];
    class subseq_code_enrl_grade/style=[cellwidth=2in];
    title "Grades for" &&course1 ;
run;
%mend table1;

%macro table2;
ods proclabel="Odds of Success" ;
proc print noobs label data=stats (keep=_RROR_ XL_RROR XU_RROR);
   var XL_RROR  _RROR_  XU_RROR ;
   format XL_RROR _RROR_ XU_RROR f4.2;
   label XL_RROR="Lower Limit"  _RROR_="Odds Ratio" XU_RROR="Upper Limit";
run;
%mend table2;
```

INTERPRETATION

Interpretation of the data analysis is straight forward, e.g., if the odds ratio is greater than one then “The odds of success are significantly higher…”. Conditional comments are pieced together from variables and macro variables to present later in the one-page report.

```
data _NULL_;
set freqs (where=(substr(prereq_code_enrl_grade,1,1)="A" and
substr(subseq_code_enrl_grade,1,2)="A"));
call symput('success1',put(round(pct_row,1),2.0));
run;

data _NULL_;
set freqs (where=(substr(prereq_code_enrl_grade,1,1)="D" and
substr(subseq_code_enrl_grade,1,2)="A"));
call symput('success2',put(round(pct_row,1),2.0));
run;
```

```
data _NULL_;
set freqs  ;
retain enrolled 0;
enrolled+count;
call symput('students',put(enrolled,comma6.0));
run;

data _NULL_;
set stats;
call symput('odds',put(round(_RROR_,1),2.0));
%global comm2a ;
low=XL_RROR;;

%let comm2=
 " For this pair of courses, the odds ratio shown in Table 2 is about &odds..** In other
words, the odds of success in &course2 &comm2a .";

%macro comm2a(low);
%if &low=1.0 %then
%let comm2a= are the same for students earning a at least a C in the prerequisite course
;
%else %if  &low <1.0 %then
%let comm2a=  are significantly lower for students earning at least a C in the prerequisite
course ;
%else %if &low >1.0 %then
%let comm2a=  are significantly higher for students earning at least a C in the prerequisite
course ;
%mend comm2a;

%let comm3=
"The purpose of this report is to provide you with evidence to support your research
hypotheses with regard to teaching improvements.  In particular, the performance of students
in this pair of prerequisite/subsequent courses are analyzed to determine whether a
satisfactory grade in a prerequisite course significantly improves the odds of success in a
subsequent course. Course grades for &students students were analyzed as a two by two table
in which the row variable was the last grade earned in &course1 and the column variable was
the first grade earned in &course2.. Table 1 and the accompanying bar chart show that
&success1 percent of the students who successfully completed &course1 subsequently passed
&course2..  Compare this number with the &success2 percent of the students who did not earn
satisfactory grades in &course1 and subsequently completed &course2.. ";
```

GRAPHICS

Graphs and charts are a good way to summarize the results of data analyses in a report and ODS LAYOUT allows us to put lots of these anywhere on a page. This scenario doesn't make full use of all the capabilities of the ODS LAYOUT feature of ODS PRINTER. The following barebones chart is included only as an example.

```
%macro chart1;
ods proclabel="Row Percentage Chart";
goptions  htext=3  fontres=presentation   ;
axis1 value=none label=none /*label=(&&COURSE1) value=( "A,B,C  "     "D,F,W*" )*/;
axis2 value=none minor=none major=none label=none ;
axis3 value=none label=none ;
pattern1 color=cx639ACE;
pattern2 color=cxE7E3E7;
title1;title2;
proc gchart data=freqs;
   hbar prereq_code_enrl_grade /
     sumvar=pct_row nostats width=6 noaxis sumlabel="Column Percent" type=sum
     subgroup=subseq_code_enrl_grade g100
     nolegend  gaxis=axis3 raxis=axis2 maxis=axis1 noframe  ;
run;
quit;
%mend chart1;
```

PRESENTATION

The experimental ODS LAYOUT feature for the ODS PRINTER destination makes it possible to put the report together. Tables and graphs of various sizes can be placed anywhere on a page. The report for this scenario is pretty straight forward but shows the basic method for producing the pdf report. Any chart or graph procedure can be "macrosized" – contained in a macro, made as small as the eye can see and placed anywhere on the page. Our example uses the absolute layout approach to specify the position of each region followed by the macro name of a comment, table or chart. The one-page report is pieced together using this basic sequence of commands:

```
ods region width=6in height=.2in x=1.9in y=7.0in;
ods printer text="Table 2. The Exact 95% Confidence Interval for the Odds Ratio";
ods region width=3in height=1.4in x=2.4in y=7.2in; %table2
```

```
ods listing close;
ods escapechar="^";
ods printer pdf file="c:\A_NESUG\Reports\Scenario1.pdf" style=styles.barrettsblue;

title1  j=c c=navy '^S={preimage="C:\A_NESUG\schreyer_weblogo50.gif"}';
title2 ;
title3 j=c c=cx639ACE "Analysis of Prerequisite/Subsequent Course Scenarios";
title4 j=c c=cx639ACE  &&course1  " by " &&course2 ;

ods layout start width=8in height=10.5in ;

ods region  width=6.5in height=2.0in  x=1in y=0in ;
ods printer text= &comm3;

ods region  width=6in height=.2in  x=2.0in y=2.0in;
ods printer text="Table 1. Relative percent of successes in the subsequent course";
ods region width=6in height=2.0in x=1.2in y=2.2in; %table1
ods region width=5.6in height=2in x=1.4in y=4.in; %chart1

ods region width=6.5in height=.8in x=1.2in y=6.0in;
ods printer text= &comm2;

ods region  width=6in height=.2in  x=1.9in y=7.0in;
ods printer text="Table 2. The Exact 95% Confidence Interval for the Odds Ratio";
ods region width=3in height=1.4in x=2.4in y=7.2in; %table2

ods region width=6in height=1.2in x=1in y=8.0in;
ods printer text='   * For the purpose of this analysis grades were categorized as "A, B or
C" or "D, F or W*"
  where "W*" includes "W", "WF","WN" and  "WP" ';
ods printer text="   **The odds ratio (or cross product ratio) is the odds of passing the
subsequent course for students passing the prerequisite
  course over the odds of passing for students failing the prerequisite course" ;
ods region width=6in height=.2in x=1in y=9in;
ods printer text=
     "Source:   &sqlobs records were obtained from the transcript table on &sysdate";

ods layout end;
ods printer close;
```

The resulting report on the following page is just a draft. ODS LAYOUT allows us the flexibility of easily replacing any pieces of the report as new ones are developed

Analysis of Prerequisite/Subsequent Course Scenarios
SAS 101 by SAS 102

The purpose of this report is to provide you with evidence to support your research hypotheses with regard to teaching improvements. In particular, the performance of students in this pair of prerequisite/subsequent courses are analyzed to determine whether a satisfactory grade in a prerequisite course significantly improves the odds of success in a subsequent course. Course grades for 597 students are analyzed as a two by two table in which the row variable was the last grade earned in SAS 101 in 200001FA and the column variable was the first grade earned in SAS 102. Table 1 and the accompanying bar chart show that 96 percent of the students who successfully completed SAS 101 subsequently passed SAS 102. Compare this number with the 75 percent of the students who did not earn satisfactory grades in SAS 101 and subsequently completed SAS 102.

Table 1. Relative percent of successes in the subsequent course

		SAS 102		
		A,B or C	D,F or W*	Total
SAS 101				
A,B or C	Number	521	20	541
	% of row	**96.30**	**3.70**	**100.00**
D,F or W*	Number	42	14	56
	% of row	**75.00**	**25.00**	**100.00**

For the course pair, the odds ratio shown in Table 2 is about 9. In other words, the odds of success in SAS 102 are significantly higher for students earning at least a C in the prerequisite course. The width of the interval shows that this measure is not very precise.

Table 2. The Exact 95% Confidence Interval for the Odds Ratio

Lower Limit	Odds Ratio	Upper Limit
3.75	8.68	19.5

* For the purpose of this analysis grades were categorized as "A, B or C" or "D, F or W*" where "W*" includes "W", "WF","WN" and "WP"
**The odds ratio (or cross product ratio) is the odds of passing the subsequent course for students passing the prerequisite course over the odds of passing for students failing the prerequisite course

Source: This report includes grades obtained from the transcript table on 03AUG04

CONCLUSION

It turns out that many of our analyses fit a certain mold. If we can reframe a typical research question to take advantage of frequently used SAS procedures, we can use macros and the ODS LAYOUT feature for ODS PRINTER to automate the interpretation and reporting of our data analyses. This paper shows an example using a simple two by two table.

FUTURE WORK

The "Prerequisite/Subsequent Course" scenario shows us that it is possible to develop a menu-driven application with data from a data warehouse and customized reports composed of preprogrammed modules for a variety of scenarios: "Student Performance", "Mid-Semester Review", "Program Evaluation", and "Retention and Progression". We plan to develop several of these scenarios and make them available on the intranet using SAS/IntrNet.

REFERENCES

Schellenberger, Brian T., (2003), "ODS LAYOUT Arranging ODS Output as You See Fit", SAS® Institute Inc., Cary NC.

ACKNOWLEDGEMENTS

SAS® is a registered trademark of SAS Institute, Inc.

CONTACT INFORMATION

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

Sue Cross
The Pennsylvania State University
301 Rider Bldg II
University Park PA 16802
(814) 863-2598

slc@psu.edu

Tim Leso
The Pennsylvania State University
148 CLRC
Altoona PA 16601
(814) 949-5287

t1l@psu.edu