CHEKOUT: A SAS® Program to Screen for Outliers

James Handsfield, Centers for Disease Control and Prevention, Atlanta, GA

ABSTRACT

At various times, most researchers are faced with the need to screen for outliers. Many statistical programs will do this with more or less ease. SAS has two PROCs which will calculate several outlier statistics. They can produce a report and output data set which require further manipulation to use. This data manipulation opens the door of opportunity for program errors which might be difficult to detect. The more statisticians can rely on known code, the more comfortable they can be with the outcome. CHEKOUT uses several macros to write the value detected as an outlier and its associated RSTUDENT statistic in the output data set from PROC GLM to a text file, OUTLIER.SAS. Also written to this file is the default option to convert the value to missing, and the file is printed as an outlier screening report. OUTLIER.SAS is easily reviewed with a text editor or in SAS , modified if necessary, and used with an %INCLUDE statement in a subsequent DATA step to remove or convert the outliers in an analysis program. Users of CHEKOUT should have knowledge of multivariable statistics and some experience with SAS macro language.

INTRODUCTION

SAS/STAT® software has two PROCs which will calculate several outlier statistics. CHEKOUT is a utility that produces a report which is easy to read and interpret, and an output file which can be easily edited using either the SAS Program Editor or any text editor. The output file can be used by including it in a data step in a subsequent analysis program.

OUTLIERS

While it is not in the scope of this paper to have an in depth statistical discussion about outliers, it will be useful to have several points in mind. There are several ways outliers may be defined depending on the type of data, whether there are known or estimated distributions, or natural constraints. In general, an outlier is an observation which is not consistent with the data, has an undue influence on the distribution of the data, or is an implausible result. An implausible result may be biologically unrealistic, such as a systolic blood pressure reading of 412. In this case, it may be that the four and the one were transposed and the correct reading is 142. Another implausible result is a value over 100 when the value is a percent.

The ideal treatment for outliers is to determine what a value should be and restore it. Unfortunately that is often not possible, and then the usual procedure is to convert the value to missing. Both of these options invite programming errors which may go completely undetected in the subsequent analysis. The fewer program edits required, the fewer errors of this type will be made.

TLI DATA

For demonstration purposes, I use T-lymphocyte immunophenotyping data collected by the Centers for Disease Control and Prevention (CDC) Public Health Practice Program Office (PHPPO), Division of Laboratory Systems (DLS) Model Performance Evaluation Program (MPEP). These data were collected from laboratories participating in the April 1997 shipment. T-lymphocyte immunophenotyping is used for diagnostic and prognostic evaluation of patients with Human Immunodificiency Virus (HIV) for determining the stage of the disease and the appropriate treatment of the patient. Most data have been collected in the form of percent CD4 or other markers. Recently, however, new technology has resulted in these data being collected as absolute counts of these markers. Because the specimens shipped to participating laboratories are alloquots of whole human blood, the exact count of any of the markers is not known. In order to evaluate performance, we assume that the mean value is the correct value, and that any response which falls within the 95% confidence interval about the mean is a correct response. We screen for outliers so that no one extreme value will expand the confidence interval and remove those values from the calculation of the reference range. In this case, we are interested in outliers only for the calculation of the confidence interval. No data are removed from the final report. For the percent data, if the upper bound of the 95% confidence interval is greater than 100, it is converted to 100. There are no negative numbers in the data. Also, while laboratories receive five tubes containing blood specimens, one tube is repeated with a different sample code, so that each laboratory tests four specimens.

We chose to use the jackknife (RSTUDENT) residual statistic because it is less influenced by the distribution of the data than some other statistics. This is necessary because we are dealing with data in different formats. The CHEKOUT program is easily modified to use other screening statistics such as Cook’s Distance, either from PROC GLM which we use, or PROC REG which will produce the same statistics.

THE PROGRAM

CHEKOUT uses five macro variables to facilitate editing of the input variables to the program. MODVARS is a list of dependent variables used in the model statement of PROC GLM. GLMVARS is a list of corresponding GLM output variables. It is essential that the variable lists used in creating these macro variables correspond exactly. STATVAL is the critical value of the statistic used to evaluate the data for outliers. PATH is a macro variable which is defined earlier in either a setup program or, in our usage, in AUTOEXEC.SAS, and defines the folder where the output file will be written. PRNTFILE contains the PATH followed by the output file name.

Two macros are used to write the outlier report to the output file. %ID writes the identifier statement to the output. These identifiers are all the information necessary to identify a unique observation and value. %SHOW writes the code for the value that resulted in the RSTUDENT value, the RSTUDENT value, and the input variable made equal to missing. The first two lines of %SHOW are commented with an asterisk and are written for the screening report.

The heart of the program is PROC GLM. Since we are not interested at this point in the printed output of GLM, use the NOPRINT option. A CLASS statement identifies the variable(s) which will be used for the independent variables in the model. The MODEL statement contains the input macro variable MODVARS which resolves to the list of output variables in the data set. The OUTPUT statement defines the name of the output data set and the contents, in this case the RSTUDENT statistic, containing GLMVARS which resolves to the names of the output variables.

The output data set is sorted by laboratory (MPEPNum) and DONOR. The next DATA step initializes a counter to determine the total number of results which will be screened for outliers. A DATA _NULL_ step sets pointer controls for the output and initializes a counter for the outliers. Control variables ID, NAME1, and NAME2 are created with a LENGTH statement, and ID is initialized with a value of 0 (zero). %PRNTFILE is identified as the output file.

Two array statements set up the variables and their corresponding RSTUDENT values, once again using MODVARS and GLMVARS. Then a DO loop sets up the screening criteria itself. In the TLI data, we wanted to identify real outliers, but not be too tight as to not eliminate data which were far from the mean but still plausible, particularly with respect to count data. We chose an absolute value of three for the screening. When the absolute value of the RSTUDENT statistic is greater than three, then %ID and %SHOW are invoked and the data printed to the output file.

When each loop of the arrays is processed, ID tests for the end of a donor and laboratory respectively, and then print END statements for the output of DONOR and MPEPNum. When EndoFile is reached, the counts will be tallied and percent outliers calculated, and this information appended to the output file. Finally a routine is run to read the output file and print it.

The output file is easily viewed and edited either with Program Manager or text editor. Once edits, if any, are completed, the output file may be utilized with an %INCLUDE statement in a DATA step with the data to be modified.

SUMMARY

CHEKOUT allows users to evaluate data for outliers using one of several statistics from PROC GLM or PROC REG. The output is a text file with an extension SAS which can be easily reviewed and edited, and then included in a subsequent DATA step.

ACKNOWLEDGEMENTS

Ronald Fehd, CDC/PHPPO/DLS was of inestimable help in proof reading CHEKOUT and offering several suggestions for improving the program. John Hancock, my supervisor at CDC/PHPPO/DLS encouraged the presentation of CHEKOUT.

James Handsfield, Ph.D. (Cand.), M.P.H.
Centers for Disease Control and Prevention
Mailstop G25
4770 Buford Highway, NE
Atlanta, GA 30341
Voice: (770)488-4164
FAX: (770)488-7663
Email: jhh0@cdc.gov

APPENDIX I

```
/* CHEKOUT.SAS --  TLI OUTLIERS REPORT *
When printing, select PRINT SETUP and select
PORTRAIT, and a left margin of 0.6.  Use SAS
MonoSpace 9 font for best presentation.
/*********************************************/
OPTIONS LS=96 PS=58 PAGENO=1 NOCENTER NOFMTERR
        MPRINT;
%LET MODVARS = CD4COMB  CD8COMB
               CD4CTCMB CD8CTCMB;
%LET GLMVARS = R4 R8 R4CT R8CT;
%LET PRNTFILE = "&PATH.\TLIOUT.SAS";
%LET STATVAL = 3; /* 1 for Cook Distance      */

/*** Macro Code *******************************/
%MACRO ID(msg);
   if NOT ID then do;
      put   "IF MPEPNum = '" MPEPNum
        +B1 "' and FormNmbr = '" FormNmbr
        +B1 "' "/" and Donor = "DONOR
        +B1 "  and SAMPID = '" SAMPID
        +B1 "' then do;";
      ID = 1; ****************************; end;
   put "  *&msg.;"; *****************; %mend ID;

%MACRO SHOW(FIELD1,FIELD2);
   CALL VNAME(&FIELD1,NAME1);
   CALL VNAME(&FIELD2,NAME2);
   OUTL+1;
   put @1 "  *" @4 NAME1 @13 " = " &FIELD1
              +B1";"/
       @1 "  *" @4 NAME2 @13 " = " &FIELD2
              +B1";"/
       @1 "   " @4 NAME2 @13 " = . "
              +B1";"; ************; %mend SHOW;
/*** End Macro Code ***************************/

PROC GLM DATA=LIBRARY.SAMPLES NOPRINT;
   CLASS DONOR;
   MODEL &MODVARS. = DONOR;
   OUTPUT OUT = OLIERS RSTUDENT = &GLMVARS.;
  *OUTPUT OUT = OLIERS COOKD = &GLMVARS.;
```

```
PROC SORT DATA=OLIERS OUT=RANGE;
   BY DONOR;

DATA _NULL_;
   retain B1 -1;    /*    Pointer Control    */
   retain TOTAL 0;  /*   Initialize counters */
   retain OUTL 0;
   file &PRNTFILE;
   set RANGE end = EndoFile;
   length ID 3;  ID = 0;
   length NAME1 $8; /* Used in MACRO SHOW */
   length NAME2 $8; /* Used in MACRO SHOW */

/*  Array Processing for output variables  */
array Outlier{*} 3  &GLMVARS.;
array Score{*}   3  &MODVARS.;

   DO i = 1 TO DIM(Score);
   IF     SCORE{i} NE . then TOTAL+1;
   IF (   ABS(Outlier{i}) GE &STATVAL )
   THEN DO;
     %ID(Review:  Outlier)
     %SHOW(Outlier{i},Score{i})
          *** ABS GE &STATVAL ***********; end;
          *** Array Processing **********; end;

/**End of Record -- NO EDITS BELOW THIS LINE */

if ID  then put "END; *MPEPNum PROCESSING;";
if EndoFile then do;
   PctOut = (OUTL/TOTAL);
   FORMAT PctOut Percent8.2;
   put "*Number of Outliers = " OUTL ";"/
       "*Number of Results  = " TOTAL ";"/
       "*Percent Outliers   = " PCTOUT ";"/
       "****End of Job;";* EndoFile ......; end;

/*IUPRNCRX Utility Print **********************
  Ronald Fehd
  97Apr04
/**********************************************/
data _NULL_;
file PRINT;
do until(EndoFile);
infile &PRNTFILE end = EndoFile pad lrecl = 88;
  input @1 Line $char88.;
  put   @1 Line $char88.;
end;
stop;
run;
```

APPENDIX II

```
/*TLIOUT.SAS:  Output from CHEKOUT.SAS      */
IF MPEPNum = '01921' and FormNmbr = '01' and
   Donor = 10  and SampID = 'A3' then do;
  *REVIEW:  OUTLIER;
  *R4         = -8.868599914;
  *CD4COMB    = 22;
   CD4COMB    = .;
END; *MPEPNum PROCESSING;
IF MPEPNum = '01939' and FormNmbr = '01' and
   Donor = 10  and SampID = 'D1' then do;
  *REVIEW:  OUTLIER;
  *R4CT       = 3.2918341362;
  *CD4CTCMB   = 1745;
   CD4CTCMB   = .;
END; *MPEPNum PROCESSING;
IF MPEPNum = '10031' and FormNmbr = '01' and
   Donor = 12  and SampID = 'B5' then do;
  *REVIEW:  OUTLIER;
  *R8         = 3.8109421572;
  *CD8COMB    = 41;
   CD8COMB    = .;
END; *MPEPNum PROCESSING;
IF MPEPNum = '10240' and FormNmbr = '01' and
   Donor = 20  and SampID = 'H1' then do;
  *REVIEW:  OUTLIER;
  *R8CT       = 5.4368312693;
  *CD8CTCMB   = 2852;
   CD8CTCMB   = .;
END; *MPEPNum PROCESSING;
IF MPEPNum = '10240' and FormNmbr = '01' and
   Donor = 24  and SampID = 'H2' then do;
  *REVIEW:  OUTLIER;
  *R4CT       = 3.1225689674;
  *CD4CTCMB   = 1855;
   CD4CTCMB   = .;
END; *MPEPNum PROCESSING;
*Number of OUTLIERS = 102 ;
*Number of RESULTS  = 7105 ;
*Percent OUTLIERS   = 1.44% ;
****END OF JOB;
```

APPENDIX III

```
/*** Typical use in a DATA step ***************/

DATA SAMPS;
   set LIBRARY.SAMPLES;
%INCLUDE SASAUTOS(TLIOUT);
/* More data statements as needed.            */
```