Reporting Outliers in Data: A Macro to Identify and Report Distant Data Points

Howard Liang, PharmaNet/i3, Eden Prairie, MN
Jason Bishop, PharmaNet/i3, Louisa, VA

ABSTRACT

When performing statistical analysis on data, certain statistical assumptions must hold. The identification of outliers is an integral part of grooming the data for analysis. One must check to see if these outliers are valid data from the data source. Reports can be created with and without detected outliers so statisticians and researchers can best decide on appropriate statistical methods and properly interpret the analysis results.

INTRODUCTION

It is a good practice to check the input data to see if there is an outlier (an unusual observation which is far from the bulk of the data) before using a given statistical model in data analysis. For a statistical model to more accurately describe the data, outliers should be removed or otherwise be accounted for in the data first. In this paper, we will use an example to show how we can detect outlying data points and then generate a report excluding outliers.

DATA SIMULATION

Suppose we have a simple clinical trial data with a sample of 30 subjects randomly assigned to placebo and a drug X with a 1:2 randomization ratio. There is one outcome measurement on ten time points. The first measurement is a pre-dose baseline assessment which is followed by nine post-dose repeat measurements. The following code produces the simulated data:

```
data sample;
  do subject=1 to 30;

    if uniform(88)>0.66 then treatment='pl';
    else treatment='dx';

    do time=0,1,2,3,4,6,9,12,16,24;
     if time>0 then
       result=100+20*normal(77);
       /**introduce larger variability by adding a factor (2/3)
          in one observation **/
       if subject=28 and time=4 then result=result*(1+2/3);
     else
       baseline=50+20*normal(77);

     if time>0 then output;
    end;
  end;
run;
```

The output produced in the SAMPLE data set is as follows:

SUBJECT	TREATMENT	TIME	BASELINE	RESULT
1	DX	1	56.3	109.9
1	DX	2	59	89.8
:				
13	PL	6	44.3	85.2
:				
30	DX	24	37.1	139.2

Output 1. Data in the SAMPLE data set.

DETECTING OUTLIERS FROM THE DATA

Suppose we apply a simple mixed model repeated measure analysis on the simulated data. The SAS®/STAT procedure PROC MIXED provides options to output studentized residuals in the model statement (OUTP =) as an output data set.

See the following code that creates the data set PRED1:

```
proc mixed data=sample;
  class subject treatment time;
  model result=treatment time treatment*time baseline/ddfm=kr outp=PRED1
  residual solution;
  repeated time/subject=subject type=cs;
  lsmeans time*treatment/diff cl alpha=0.2 slice=time;
  ods output diffs=diff1 lsmeans=lsm1;
run;
```

The data produced by the PROC MIXED code above is stored in the work data set PRED1 and would have the following form:

SUBJECT	TREATMENT	TIME	BASELINE	RESULT	STUDENTIZED RESIDUAL
1	DX	1	56.3	109.9	0.46
:					
21	PL	1	43.4	61.2	-2.24
:					
28	DX	4	48.1	185.5	4.08
:					
30	DX	24	37.1	139.3	1.89

Output 2. Data in the PRED1 data set.

Checking standardized residuals is a good technique in the detection of outliers, as we can expect about 95% of standardized residuals to be within ±2 standard deviations. So a standardized residual larger than the absolute value of 3 can be investigated as a potential outlier, since that would only be expected to occur randomly about 0.3% of the time (probnorm(-3) x 2 = 0.0027). But the actual outlier cut-off which is acceptable often depends upon the number of observations, so this could vary among study data sets.

Statisticians are more concerned at the risk of incorrectly removing any residual so a bigger outlier cut-off point is often recommended. For our example, a clinical study with 30 subjects receiving two treatments in nine periods, there are 270 observations. By using a standard normal distribution, the probability of incorrectly removing any residual with 3σ (probnorm(-3) x 2 = 0.0027) or cutoff of 3 = $1-(1-0.0027)^{270}$ = 52% . Implementing with 4σ (probnorm(-4)x2=0.000063) or a cutoff of 4 ($1-(1-0.000063)^{270}$) would yield 1.7%.

Before running the reporting macro, we can first look at the studentized residuals vs. a predicted values plot of the PRED1 data set above.

The following SAS/GRAPH code creates the scatter plot show in Figure 1 below:

```
axis1 label=("Predicted Value") value=(h=1.5);
axis2 label=(angle=90 "Studentized Residual") minor=(n=1);
symbol1 h=3 w=2 v=circle;

proc gplot data=pred1;
  plot studentresid*pred/haxis=axis1 vaxis=axis2 vref=-2 -1 0 1 2 3 4 lvref=2
  cvref=ltgray;
run;
quit;
```

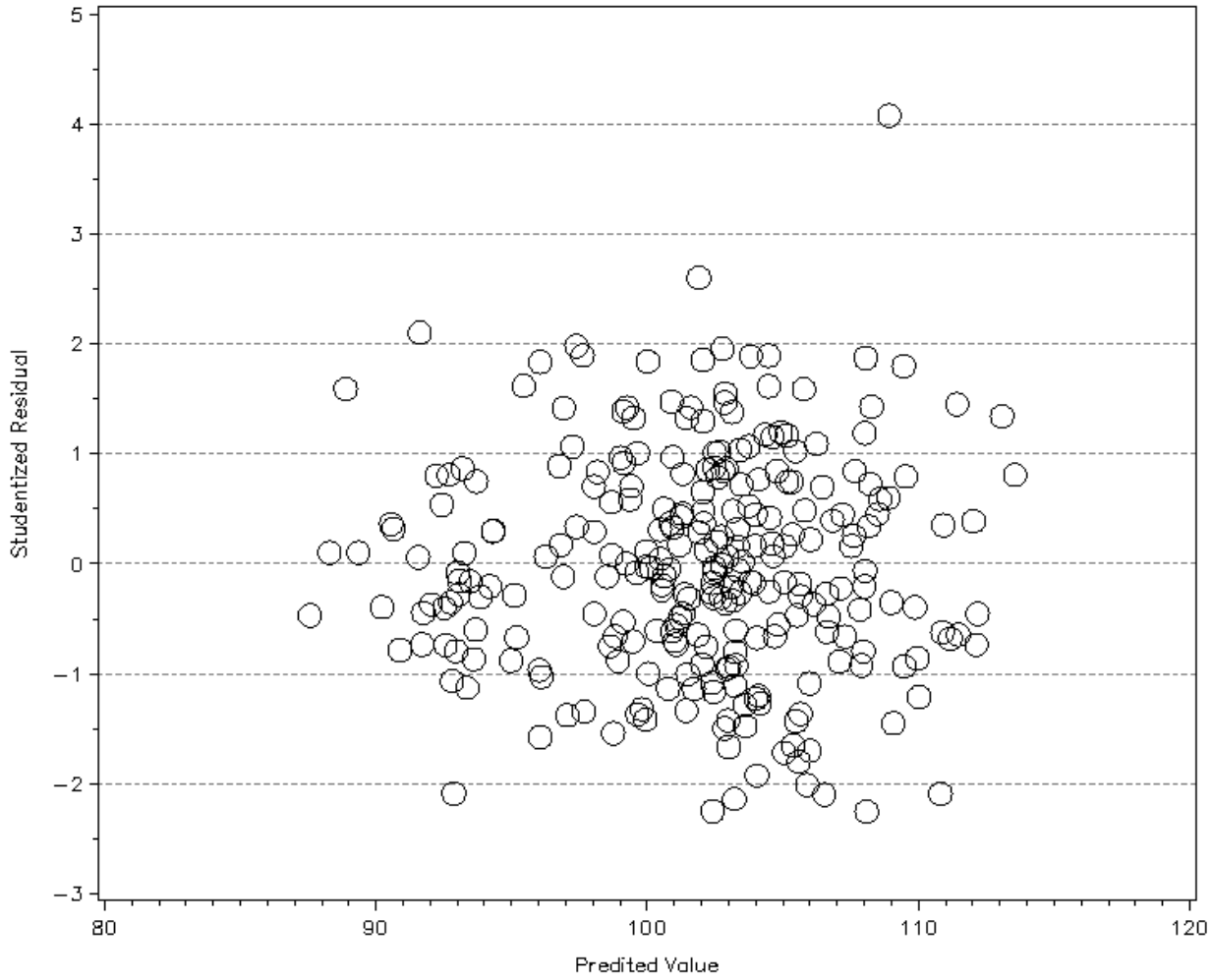

Figure 1. Predicted Values vs. Studentized Residual for data in the PRED1 data set.

This model checking plot provides a visual assessment of how the data are spread. One can easily check for extreme data points and specify the rule to detect outliers. For this sample data, an outlier cutoff-point of <-4 or >4 is chosen from the studentized residuals.

A MACRO TO REMOVE OUTLIERS

By specifying an outlier cut-off point, this macro creates an outlier data set by comparing the outlier cut-off point to studentized residual values from the output data set PRED1 generated by the previous PROC MIXED code. If outlier data exists, such values will be taken out automatically in the new report. In addition, the outliers excluded from the data analysis also are displayed in the report footnotes.

```
%macro find_outlier(sr=);

  data outlier(keep=subject treatment time);
    set pred1;
    if .<studentresid<-&sr. or studentresid>&sr.;
  run;

  %if %sysfunc(exist(outlier)) %then %do;

    proc sort data=outlier;
      by subject treatment time;
    run;

    proc sort data=sample;
      by subject treatment time;
    run;

    data sample2;
      merge sample(in=a) outlier(in=b);
      by subject treatment time;
      if a;
      if a and b then delete;
    run;

    *** model after removing outliers;
    proc mixed data = sample2;
      class subject treatment time;
      model result = treatment time treatment*time baseline / ddfm = kr;
      repeated time / subject = subject type = cs;
      lsmeans time*treatment / diff cl alpha=0.2 slice=time;
      ods output diffs=diff1b lsmeans=lsm1b;
    run;

    *** create 1 macro variable for footnoting outliers information;
    data _null_;
      set outlier end=last;
      length string $ 120;
      retain string;

      if subject ne . and not last then
        string=strip(string)||' subject '||strip(subject)||' with treatment'||
               strip(treatment)||' at time '||strip(time)||',';

      if subject=. then string=strip(string)||'.'||',';

      if last then do;
        string=strip(string)||' subject '||strip(subject)||' with treatment'||
               strip(treatment)||' at time '||strip(time);
         call symputx('string',string);
      end;
    run;

    *** create reports;
    proc report
    :
    line@3 "  (&string.)";   *** referecing the macro variable for outliers
    :                           information;
  %end;  *** checking outlier exist;

%mend find_outlier;

%find_outlier(sr=4);           *** outlier cut-off point;
```

The following output is an example of what a report would look like with outliers excluded:

Treatment			Test		Reference				
Test	Reference	Time in hour(s)	n	LSmean	n	LSmean	Difference	80% CI	P-value
DX	PL	1	19	100.23	11	103.21	-2.98	(-12.11, 6.146)	0.6750
DX	PL	2	19	100.46	11	101.73	-1.28	(-10.36, 7.808)	0.8570
DX	PL	3	19	104.11	11	100.82	3.29	(-5.796, 12.376)	0.6421
DX	PL	4	18	104.87	11	100.95	3.92	(-5.253, 13.094)	0.5834
DX	PL	6	19	105.36	11	93.168	12.20	(3.111, 21.281)	0.0858
DX	PL	9	19	100.36	11	94.122	6.24	(-2.854, 15.338)	0.3787
DX	PL	12	19	101.68	11	95.761	5.91	(-3.174, 15.004)	0.4038
DX	PL	16	19	106.84	11	104.90	1.95	(-7.141, 11.035)	0.7833
DX	PL	24	19	105.48	11	91.199	14.28	(5.200, 23.367)	0.0444

```
Abbreviations: CI = Confidence Interval; LSmean = Least Square Mean; DX = Drug X;
               PL = Placebo
Model: Absolute Values = treatment time treatment*time baseline
Note:  Outliers defined as standardized residual <-4 or >4
       Analysis excludes the following: Subject 28 with treatment DX at time 4
```

Output 3. Sample report with outliers excluded.

CONCLUSION

This macro is used to look for outliers where such a data point is defined as a residual that is set apart from the bulk of the data. When outliers exist, it provides analyses with outliers excluded, so one can ensure the assumptions one makes in the statistical model hold true. Finally, when outliers are found, it is also helpful to provide analyses twice, once with outliers excluded and again with outliers included.

CONTACT INFORMATION

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

Name: Howard Liang
Enterprise: PharmaNet/i3
Address: 12125 Technology Drive
City, State ZIP: Eden Prairie, MN 55344
Work Phone: (952) 833-7462
E-mail: HLiang@pharmanet-i3.com
Web: www.pharmanet-i3.com

Name: Jason Bishop
Enterprise: PharmaNet/i3
Address: 1787 Sentry Parkway West, Suite 300, Building 16
City, State ZIP: Blue Bell, PA 19422 USA
Work Phone: (540) 967-2109
E-mail: JBishop@pharmanet-i3.com
Web: www.pharmanet-i3.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.