PharmaSUG China 2017 - Paper 16

Outlier analysis in SAS

Focus on the programming approach

Chao Jiang, MSD, Beijing, China

Yan Liu, MSD, Beijing, China

ABSTRACT

Outlier, sometime because of noise and pollution which can confound analysis and reporting, but sometimes it contains valuable result which may impact the efficacy comparison in clinical trials. This paper presents several detection approaches used to identify the outliers and implement using multiple SAS procedures and functions. It also provides the introduction of regression model analyzing, by which to find out the effect of outliers on a model of relationships among variables, from SAS programming perspective.

INTRODUCTION

Outlier, sometime because of noise and pollution which can confound analysis and reporting, but sometimes it contains valuable result which may impact the efficacy comparison in clinical trials.
There are some reasons for the outlier, including:

- Incorrect collection
- Natural variation
- Wrong selection of sample
- Some factors impact

But in our clinical trial studies, after many times clean work by clinical operation team or data management team, wrong data become very rare in database. In my following topic I will focus on the analysis of outlier caused by some factors behind.

HOW TO IDENTIFY THE OUTLIER

Taking about the outlier, my first topic here is about how to identify the outlier. Generally speaking, there are two methods to find them in our clinical trials:
1. Clinical approach – the criteria will be given from clinical team or expert, so I will not discuss this in my following context; and
2. Statistical approach. In my study, we have two criteria a) any observation value greater than the 3rd quartile plus 1.5 times the interquartile range; and b) any observed time where the Studentized residual from the ANOVA model is > +2.
Of course, we also generated some plots to explore the outlier in the very beginning.
I want to firstly talk about why we choose above statistical approach. Maybe you know a common rule-of-thumb taught in elementary statistics about outlier is that 95% of the data in a distribution which (normally distributed) will lie within 1.96 standard deviates of the mean of the distribution. A more robust technique was proposed by Tukey. This technique is robust because it uses the quartile values instead of variance to describe the spread of the data. And, quartiles are less influenced by extreme values. This figure 1 hope will help you remember the normal distribution and the α is 5% which we prefer used in our daily work.

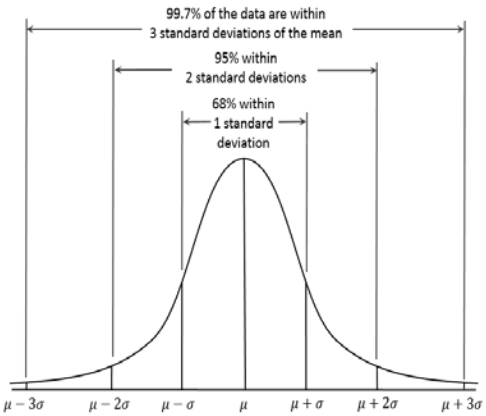

Figure 1. Normal distribution

IDENTIFY THE OUTLIER VIA PLOTS – BOX PLOT

Using figures at the very first step, helping us understanding the distribution of data and get the initial idea on further analysis direction.
Below is the code we generate the box plot.

```
data style;
  length value FillColor LineColor $30;
  Id='myid'; Value="A"; FillColor='white'; LineColor='Black'; output;
  Id='myid'; Value="B"; FillColor='grey'; LineColor='Black'; output;
run;

proc sgplot data=indata dattrmap=style;
   vbox  value / group=grpvar groupdisplay=cluster grouporder=ascending
                 category=catvar fill
                 meanattrs=(color=black symbol=squarefilled size=6)
                 medianattrs=(color=black pattern=solid thickness=2)
                 outlierattrs=(color=black symbol=circle size=6)
                 lineattrs=(color=black pattern=solid)
                 connectattrs=(color=black pattern=solid)
                 whiskerattrs=(color=black pattern=solid)
                 attrid=myid;
   yaxis discreteorder=data display=(nolabel)  ;
   xaxis label=" ";
run;
```

Firstly we created a dataset named style, which defines the box's style, including a) the ID name, which will be used in parameter ATTRID in SGPLOT procedure – in my example the ID is X; b) the name from different group - in my example there are A & B; c) the color filled the box – in my example the color are white and grey separately; d) the line' s color – for the two groups are all black. Next draw the box plot via SGPLOT procedure. The STYLE dataset we have defined before is used in DATTRMAP parameter. You can find there are many attribution parameters here to control the appearance of the box and lines. For more details please refer to SAS help. Looking at this figure, you may predict that these dots far from the 1st interquartile value maybe the outlier value. Please pay attention to the GROUP & CATEGORY parameters, the GROUP here used to distinguish different visual attributes in plot elements automatically, and CATEGORY here to create a box plot for each category – in my example I have set SEX as CATEGROY value, so there are two sets of plots, and each set including two groups – Group A and Group B. Figure 2 is the result from above code.

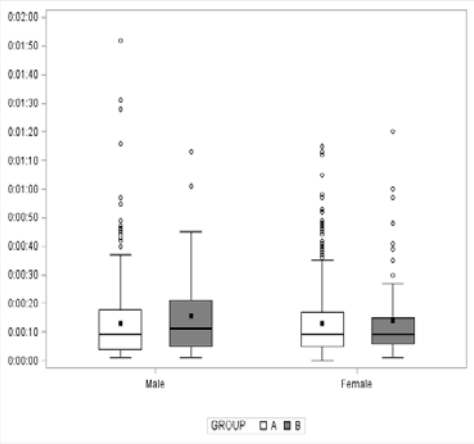

Figure 2. Box plot result

IDENTIFY THE OUTLIER VIA PLOTS – SCATTER PLOT

Below is the code we generate the scatter plot.

```
data style;
  length value markercolor markersymbol $30;
  Id='myid'; Value="A"; markercolor='grey'; markersymbol='circle'; output;
  Id='myid'; Value="B"; markercolor='black'; markersymbol='circlefilled'; output;
run;

proc sgplot data=indata dattrmap=style;
   scatter x=xvar y=yvar / group=grpvar groupdisplay=cluster
                           attrid=x;
   yaxis label="ylabel";
   xaxis label="xlabel";
   keylegend / location=outside position=top
run;
```

For scatter plot, same as box plot, firstly we defined the style dataset, but different to above, here we used MAKERSYMBOL to control the dots' appearance. In my example, one group is circle but without filling any color and another group is filling-color-circle. And then we draw the plot via SGPLOT procedure. From the plot (figure 3) we can predict that seems group A has more outliers than group B.

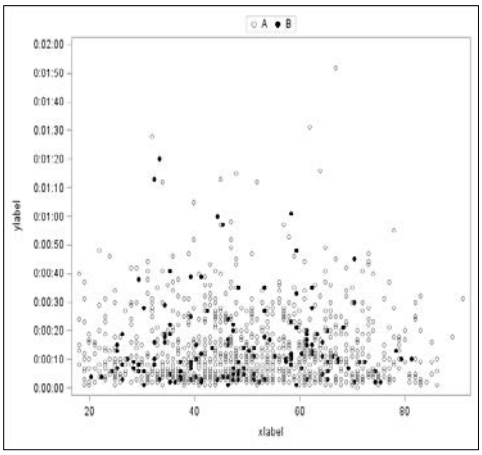

Figure 3. Scatter plot result

IDENTIFY THE OUTLIER VIA PLOTS – STATISTICAL APPROACHES 1

The first statistical approach to identify the outlier is criteria that any observation value greater than the 3rd quartile plus 1.5 times the interquartile range. Here we used the UNIVARIATE procedure (the code displayed as below).

```
proc univariate data = indata noprint;
  class byvar1 byvar2;
  var var;
  output out = outdata
     q1= q1
     q3= q3
     qrange=qrange ;
run;

If value > q3+1.5*qrange >. then set the flag = 'Y'
```

All the parameters here in Italy front can be changed according to your study. From the outdata dataset, we can obtain the q3 value and the interquartile range. So any observed value if greater then q3+1.5qrange we will flag that as Y means which met our first criteria.

IDENTIFY THE OUTLIER VIA PLOTS – STATISTICAL APPROACHES 2

The second statistical approach to identify the outlier is criteria that any observed time where the Studentized residual from the ANOVA model is > +2. Here we used the GLM procedure (the code displayed below).

```
proc glm data = indata noprint;
  class classvar1 classvar2 classvarn ;
  model value = classvar1 classvar2 classvarn;
        output out = anova predicted=pred residual = res  student = student;
quit;

If student > 2 then set flag = 'Y'
```

Same as last one, all the parameters here in Italy front are changeable depend on your case. But here there are some tips: a) variables in CLASS statement will be used in the model, that means all the independent effects variables should appears in model statement, and the CLASS statement must appear before the MODEL statement. And classification variables can be either character or numeric. For MODEL statement, you can only specify only one. The code displayed here is the simplest usage, you can add any options if needed. We can obtain the Studentized residual value from the output dataset, and for any observation's Studentized residual value greater than 2 will be flagged – should pay attention that this is group-by-group value not patient-by-patient, means all records belong to one group shared the same value from the model.

EVALUATION POTENTIALLY FACTORS

So far, we have flagged all outliers according to our criteria set before. Next step, we want to evaluate the potential factors – to figure out what cause these outliers. The method we used is regression model, but here we need discuss category / numeric variables separately – there are different solutions.

EVALUATION POTENTIALLY FACTORS - FOR CATEGORY EXPLANATORY VARIABLE

For category variables we used LOGISTIC procedure. Below is the code we used.

```
proc logistic data = indata  ;
  class classvar1 classvar2 var1(ref="ref1") var2(ref="ref2")/param=glm;
  model flagvar = classvar1 classvar2 var1 var2  var1* var2/CLPARM=wald;
    oddsratio var1 ;
    slice var1*var2 / sliceby = var1 diff oddsratio cl ;
    ods output SliceDiffs = Slicediff ;
run;
```

There are some tips I want to mention here: a) the CLASS statement must precede the MODEL statement and all the variables in CLASS statement will be used as explanatory variables in the below MODEL statement. In my example, the class variable including classvar1, classvar2, var1 and var2 – there are all in MODEL effects variables; b) the CLPARM option here specify the method to generate the CI, and in my study we chose WALD; c) the ODDSRATIO

statement produces odds ratios for variable which you specified following, and that variable can be continuous or classification – in my example we generated the odd ratio for var1; d) the SLICE statement here provides a general mechanism for performing a partitioned analysis of the LS-means for an interaction, which is also known as simple effects – there will be more details later. But before that, I want to mention the convergence problem. You can specify the convergence criteria in MODEL statement – *ABSFCONV*, *FCONV*, *GCONV*, or *XCONV*. And the convergence info will be print in output result, and also you can find this WARNING (Figure 4) in log which indicates you model need to reverse and the solution is decreasing the classification variables. For example in my previous code, you may remove classvar2 from MODEL statement and check the log or result again, and repeat this step until the model is fine.

```
WARNING: There is possibly a quasi-complete separation of data points. The maximum likelihood
         estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are
         based on the last maximum likelihood iteration. Validity of the model fit is
         questionable.
```

Figure 4. Not convergence WARNING in log

Below shows the parameter estimates results and the 95% CI for Wald. The second column from left displays the different level for each parameter. For example, for BYVAR1C there are two levels – 20 and 40. And the p-values are from Type3 Analysis of Effects page.

Parameter Estimates and Wald Confidence Intervals

Parameter			Estimate	95% Confidence Limits	
Intercept			-1.2097	-2.5018	0.0823
BYVAR1C	20		-0.7656	-1.6550	0.1238
BYVAR1C	40		-1.1407	-2.0776	-0.2039
BYVAR2C	9		-0.9745	-1.5411	-0.4079
SEXF	Female		-0.7013	-2.1274	0.7247
TYPE	D		-0.7584	-2.1749	0.6581
TYPE	J		1.0404	0.0706	2.0103
TYPE	K		0.7853	-0.2467	1.8173
TYPE	L		-9.7099	-487.4	468.0
TYPE	N		-0.3878	-1.8342	1.0585
SEXF*TYPE	Female	D	1.0928	-1.2244	3.4100
SEXF*TYPE	Female	J	-0.1246	-1.7815	1.5324
SEXF*TYPE	Female	K	-8.9681	-155.1	137.2
SEXF*TYPE	Female	L	10.9636	-466.7	488.7
SEXF*TYPE	Female	N	1.8615	-0.2214	3.9444

Figure 5. parameter estimates and 95% CI for Wald

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
BYVAR1C	2	6.1756	0.0456
BYVAR2C	1	11.3639	0.0007
SEXF	1	0.0000	0.9981
TYPE	5	8.7243	0.1206
SEXF*TYPE	5	5.9615	0.3100

Figure 6. P-values from Type3 Analysis of Effects

For SLICE statement, this provides the p-value, odd ratio and CI for each interaction classification as we specified these results in SLICE statement. Figure 7 shows where this info we can obtain from output, but here I just display the slice result for Female, for male the results will be printed in another output page.

Simple Differences of SEXF*TYPE Least Squares Means

Slice	TYPE	_TYPE	Estimate	Standard Error	z Value	Pr > |z|	Alpha	Lower	Upper	Odds Ratio	Lower Confidence Limit for Odds Ratio	Upper Confidence Limit for Odds Ratio
SEXF Female	D	J	-0.5814	0.8025	-0.72	0.4688	0.05	-2.1544	0.9915	0.559	0.116	2.695
SEXF Female	D	K	8.5173	74.5721	0.11	0.9091	0.05	-137.64	154.68	>999.999	<0.001	>999.999
SEXF Female	D	L	-0.9193	0.7554	-1.22	0.2236	0.05	-2.3999	0.5613	0.399	0.091	1.753
SEXF Female	D	N	-1.1392	0.8704	-1.31	0.1906	0.05	-2.8451	0.5667	0.320	0.058	1.763
SEXF Female	D	Z	0.3344	0.9353	0.36	0.7207	0.05	-1.4987	2.1675	1.397	0.223	8.738
SEXF Female	J	K	9.0987	74.5694	0.12	0.9029	0.05	-137.05	155.25	>999.999	<0.001	>999.999
SEXF Female	J	L	-0.3379	0.4113	-0.82	0.4113	0.05	-1.1440	0.4682	0.713	0.319	1.597
SEXF Female	J	N	-0.5578	0.5998	-0.93	0.3524	0.05	-1.7334	0.6179	0.572	0.177	1.855
SEXF	J	Z	0.9159	0.6896	1.33	0.1841	0.05	-0.4357	2.2674	2.499	0.647	9.654

Figure 7. SLICE statement outputs for Female level

EVALUATION POTENTIALLY FACTORS - FOR CONTINUOUS EXPLANATORY VARIABLE

Above we discussed the regression model for category in explanatory variables side. Next I want to share some tips on continuous. The code (below) seems very same as category's, but only one tips, the odd ratio we should obtain from the Odds Ratio Estimates page. Figure 8 shows this results for my example. And we get the p-value from Analysis of maximum likelihood esitmates page. Figure 9 shows this.

```
proc logistic data = indata ;
  class classvar (ref="ref1")/param=ref;
  model flagvar = classvar var/CLPARM=wald;
run;
```

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
BYVAR1 20 vs 160	0.403	0.169	0.965
BYVAR1 40 vs 160	0.336	0.137	0.827
BYVAR2 9 vs 20	0.336	0.191	0.591
AGE	1.034	1.019	1.050

Figure 8. Odds Ratio and 95% Wald CI

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.5377	0.6007	17.8464	<.0001
BYVAR1	20	1	-0.9078	0.4449	4.1630	0.0413
BYVAR1	40	1	-1.0899	0.4590	5.6378	0.0176
BYVAR2	9	1	-1.0901	0.2881	14.3174	0.0002
AGE		1	0.0338	0.00778	18.8439	<.0001

Figure 9. P-value for continuous explanatory variables

LESSON LEARNT AND TIPS

- Always consult with statistician if any problem as they are experts in this topic and decide what contents

they want to support their paper

- Check the SAS help for each statement in model
- High recommend that the QC side use another software to validate the results (R, excel)

REFERENCES

SAS Help and Documentation: http://support.sas.com/documentation/onlinedoc/91pdf/index_913.html

CONTACT INFORMATION

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

Name: Chao Jiang
Enterprise: MSD R&D China
Address: 1-13F, Building 21 Rongda Road, Wangjing R&D Base, Zhongguancun Electronic Zone West Zone, Chaoyang District
City, State ZIP: Beijing 100012, China
Work Phone: 1058609387
E-mail: chao.jiang@merck.com

Name: Yan Liu
Enterprise: MSD R&D China
Address: 1-13F, Building 21 Rongda Road, Wangjing R&D Base, Zhongguancun Electronic Zone West Zone, Chaoyang District
City, State ZIP: Beijing 100012, China
Work Phone: 1058609402
E-mail: yan.liu14@merck.com