Paper SP06-2009

DETECTION OF MULTIPLE OUTLIERS IN UNIVARIATE DATA SETS

Marek K. Solak, PhD
Schering-Plough Research Institute, Summit, NJ

ABSTRACT
A number of methods are available to detect outliers in univariate data sets. Most of these tests are designed to handle one outlier at a time. As soon as an outlier is found it is removed from the data set and the process is repeated until no more outliers are detected. Grubbs (1950, 1969) and Dixon (1953) tests can handle, in some cases, more than one outlier at a time. However, in general, when multiple outliers are present masking phenomenon (an outlier is not detected, due to presence of other outliers) may prevent outlier detection. PROC ROBUSTREG appears as a useful tool to evaluate multiple outliers. It has four types of estimation available (M, LTS, S, MM). Performance of PROC ROBUSTREG will be compared with sequential application of Grubbs test, 3 sigma and Weisberg t-test. SAS® macros to implement multiple outliers testing will be presented as well.

KEYWORDS
Grubbs test, masking phenomenon, outlier, PROC ROBUSTREG, 3 sigma test, t-distribution, Weisberg t-test.

INTRODUCTION

Two sided Grubbs test (Grubbs 1950) is often used to evaluate measurements, coming from a normal distribution of size n, which are suspiciously far from the main body of the data. For a two-sided Grubbs' test, the test statistic is defined as:

$$G = \frac{|y_O - \bar{y}|}{s} \tag{1}$$

with $\bar{y}$ and *s* denoting the sample mean and standard deviation, respectively, calculated with the suspected outlier included. The critical value of the Grubbs' test is calculated as

$$C = \frac{(n-1)}{\sqrt{n}} \sqrt{\frac{t^2_{(\alpha/2n,n-2)}}{n-2+t^2_{\alpha/2n,n-2}}} \tag{2}$$

where $t_{(\alpha/2n,n-2)}$ denotes the critical value of the t-distribution with (*n*-2) degrees of freedom and a significance level of $\alpha/2n$. If $G \geq C$, then the suspected measurement is confirmed as an outlier, cf. (Grubbs 1950, p. 29).

Critical Grubbs values can be easily calculated using SAS® macro:

```
%macro grubbs_crit(alpha=0.05, num=, ds=);
   data &ds;
      t2=tinv(&alpha/(2*&num), &num - 2);
      gcrit2=((&num-1)/sqrt(&num))*sqrt(t2*t2/(&num-2+t2*t2));
      label gcrit2='Critical (95%) Two-sided Grubbs Multiplier';
   run;
%mend grubbs_crit;

%grubbs_crit(num=<num>, ds=<ds>);
```

– macro call with default significance level of 0.05, alpha – significance level, num – a number of observations in a data set (ds), ds – an input data set.

Grubbs test appears to be a very useful tool in single outlier detection. In case of multiple suspect observations present, like a few measurements lumped together away from the main body of data, sequential application of Grubbs test may not detect (due to masking effect) outliers correctly. The phenomenon of masking (the presence of two (or more) outliers may prevent detection of one (or more) of them) is discussed in the literature; for references and discussion (Beckman and Cook 1983). Sequential (the smallest (largest) measurement is tested first, then the

next in line and so on) application of the Grubbs test may result, in some cases, in false classification of the first (or even several consecutive) suspect results as non-outliers. Grubbs provided statistics (Grubbs 1950) and critical values to consider simultaneous evaluation of two suspect observations for samples of size 4-20.

In case of a doubt, when applying sequentially Grubbs test, SAS® Proc Robustreg may be useful.

TEST SAMPLE

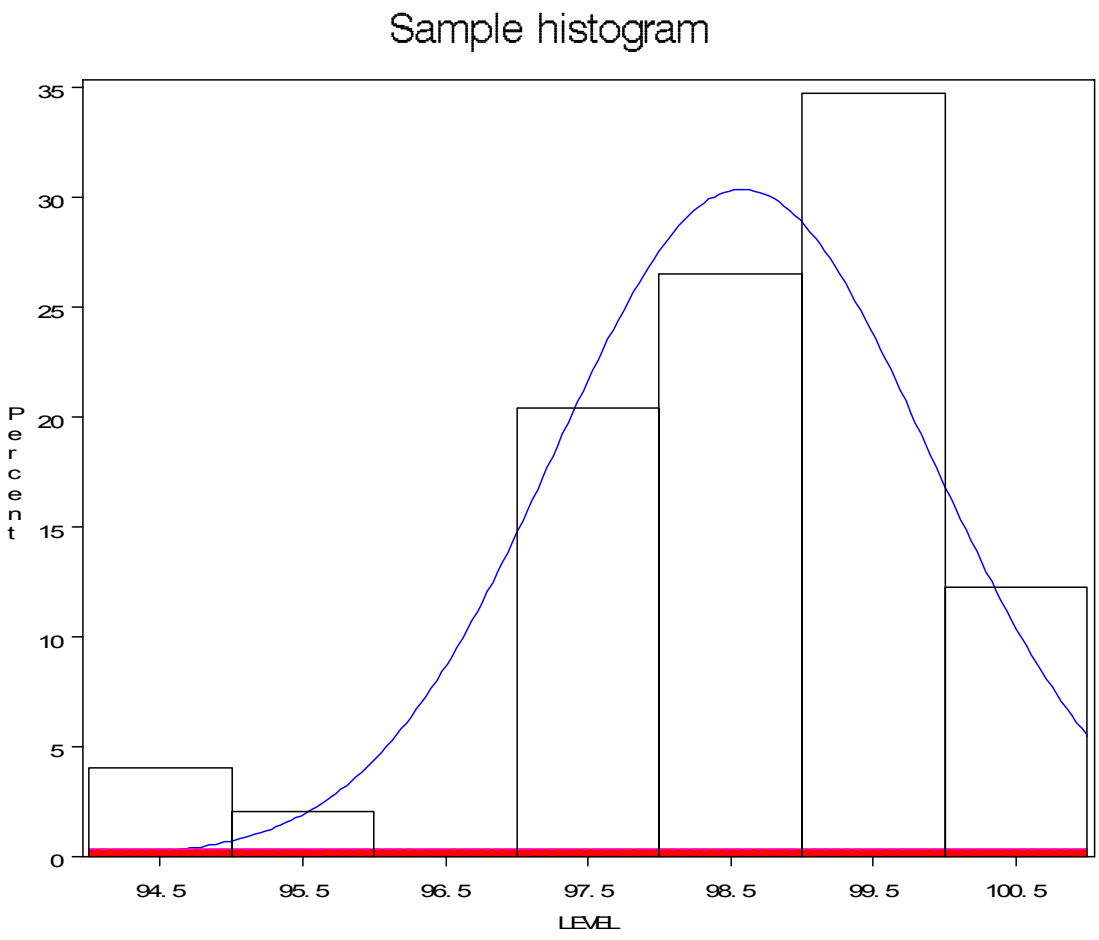
Consider a (ordered) set of measurements (49 observations, mean = 98.58, STD = 1.31)

94.8	97.6	98.2	98.4	99.2	99.4	99.8
94.8	97.7	98.2	98.5	99.2	99.4	100
95.1	97.8	98.2	98.6	99.2	99.5	100.1
97	97.8	98.3	98.7	99.3	99.5	100.2
97	97.9	98.4	98.9	99.3	99.5	100.3
97.1	97.9	98.4	99.2	99.4	99.6	100.3
97.3	98.2	98.4	99.2	99.4	99.7	100.5

(3)

Grubbs test (Grubbs 1950) shall be applied to normally distributed data sets of sizes greater than 6 and less than 50. A number of criteria exist to evaluate distribution normality for instance, -2 ≤ skewness ≤ 2 and -2 ≤ kurtosis ≤ 2. For the above sample skewness = -1.23 and kurtosis = 1.84. Skewness and kurtosis can be calculated using PROC UNIVARIATE. The same procedure may provide a histogram of the sample distribution.

```
proc univariate data=Grubbs_test noprint;
    var level;
    histogram level/exp(fill l=3) cfill=red normal(noprint);
    output out=out_Gtest std=std mean=mn skewness=skw kurtosis=kurt;
    title1 "Sample histogram";
run;
```



The sample can be viewed as a vector of 49 measurements with values $y_i = \bar{y} + e_i$, i = 1, 2, … , 49, $\bar{y}$ - average of all 49 measurements, e_i – unknown error of normal distribution. Now we may use the usual linear model (Chen 2002).

$$\mathbf{y} = \mathbf{X}\theta + \mathbf{e},\ \mathbf{X} = [1, 1, \ldots,1]^T,\ \theta = \bar{y} \qquad (4)$$

SAS®/STAT ROBUSTREG PROCEDURE

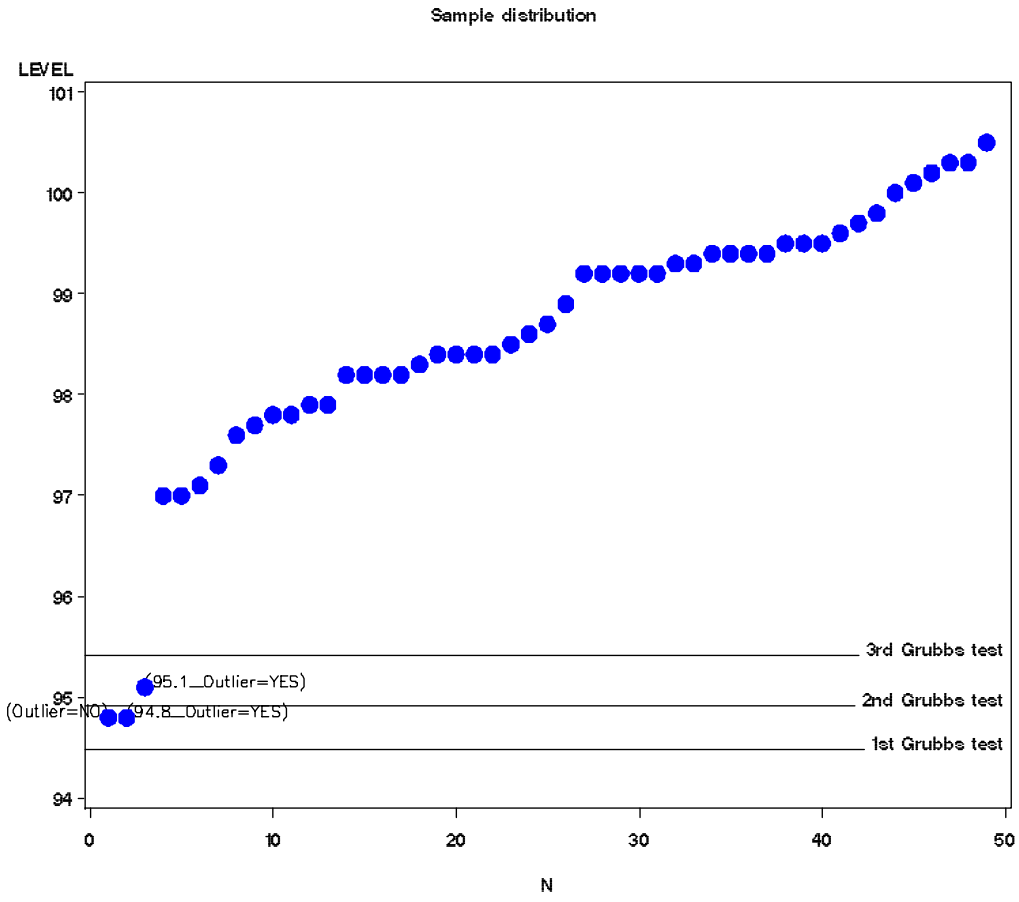
The ROBUSTREG procedure is experimental one in SAS/STAT® version 9. This procedure implements commonly used regression techniques (Chen 2002) including M, LTS, S and MM estimation. The following statements invoke the ROBUSTREG procedure with MM estimation (in case of model (4) leverage and rho options are not feasible).

```
proc robustreg data=robust_reg method=MM;
    model level=t/ diagnostics(all);
    output out=diag_out r=resid sr=stdres outlier=otlr leverage=lvr rho=r;
run;
```

Input data set, robust_reg, shall include variable **t: t =1**; to apply model (4).

GRUBBS TEST VERSUS ROBUSTREG PROCEDURE

If Grubbs (double-sided with 95% confidence level) test is applied to the first suspicious observation (=94.8; (3)) this observation is not classified as an outlier. After the first observation is removed from the data set, the next observation (=94.8, (3)) is confirmed by Grubbs as an outlier, as well as the third one, when repeated Grubbs test is applied to 95.1 (both 94.8 and 94.8 observations are removed, (3)). Grubbs test failure to classify the first observation as an outlier can be attributed to the masking effect of two additional outliers (94.8 and 95.1, (3)). Masking effect may occur when several outliers are clustered together away from the main body of data. The phenomenon of masking (the presence of two (or more) outliers may prevent detection of one (or more) of them) is discussed in the literature; (Beckman and Cook 1983). Sequential application of the Grubbs test may result, in some cases, in false classification of the first (or even several consecutive) suspect results as non-outliers.

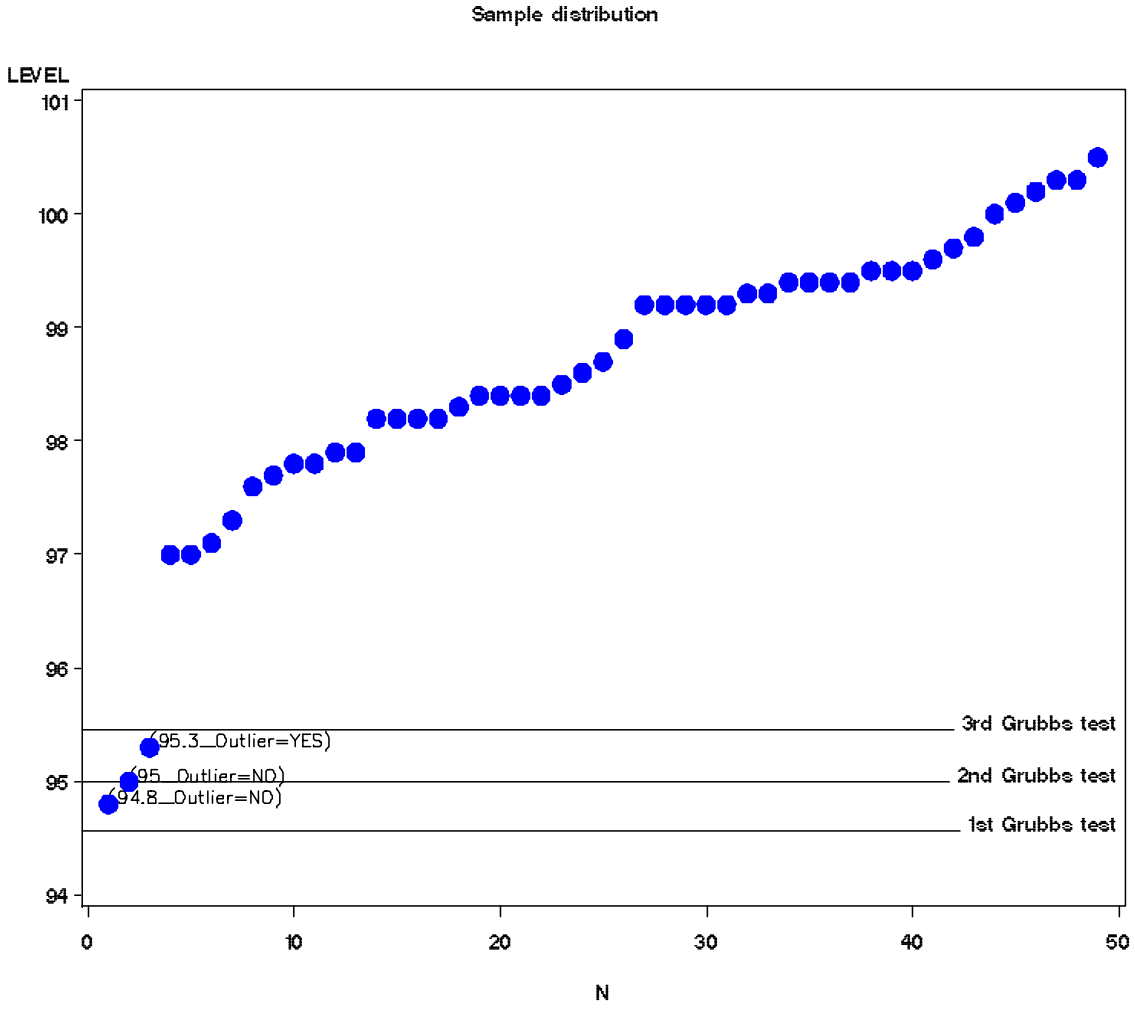


Standarized PROC ROBUSTREG residuals (cutoff = 3.0000)			
Method = M	Method = LTS	Method = S	Method = MM
-3.7735	-3.6377	-3.5731	-3.5575
-3.7735	-3.6377	-3.5731	-3.5575
-3.4896	-3.3654	-3.3088	-3.2936

More detailed results are provided below using PROC ROBUSTREG (Method = M).

Level	Robustreg Standarized Residual	Robustreg Outlier	Sample Mean	Sample Std	Grubbs Critical Value	Grubbs Outlier	Grubbs Statistics
94.8	-3.773477	1	98.579592	1.3135479	3.1201277	0	2.8773916
94.8	-3.773477	1	98.658333	1.2049249	3.1117965	1	3.202136
95.1	-3.489564	1	98.740426	1.0737297	3.1032431	1	3.3904488

If the second and the third measurements are slightly different (94.8, 95.0, 95.3) sequential application of Grubbs test misses two first suspect observations, namely 94.8 and 95.0 are not classified as outliers. Robustreg (Grubbs) outlier (=1) – an outlier is confirmed (=0) – a suspect result is not an outlier.

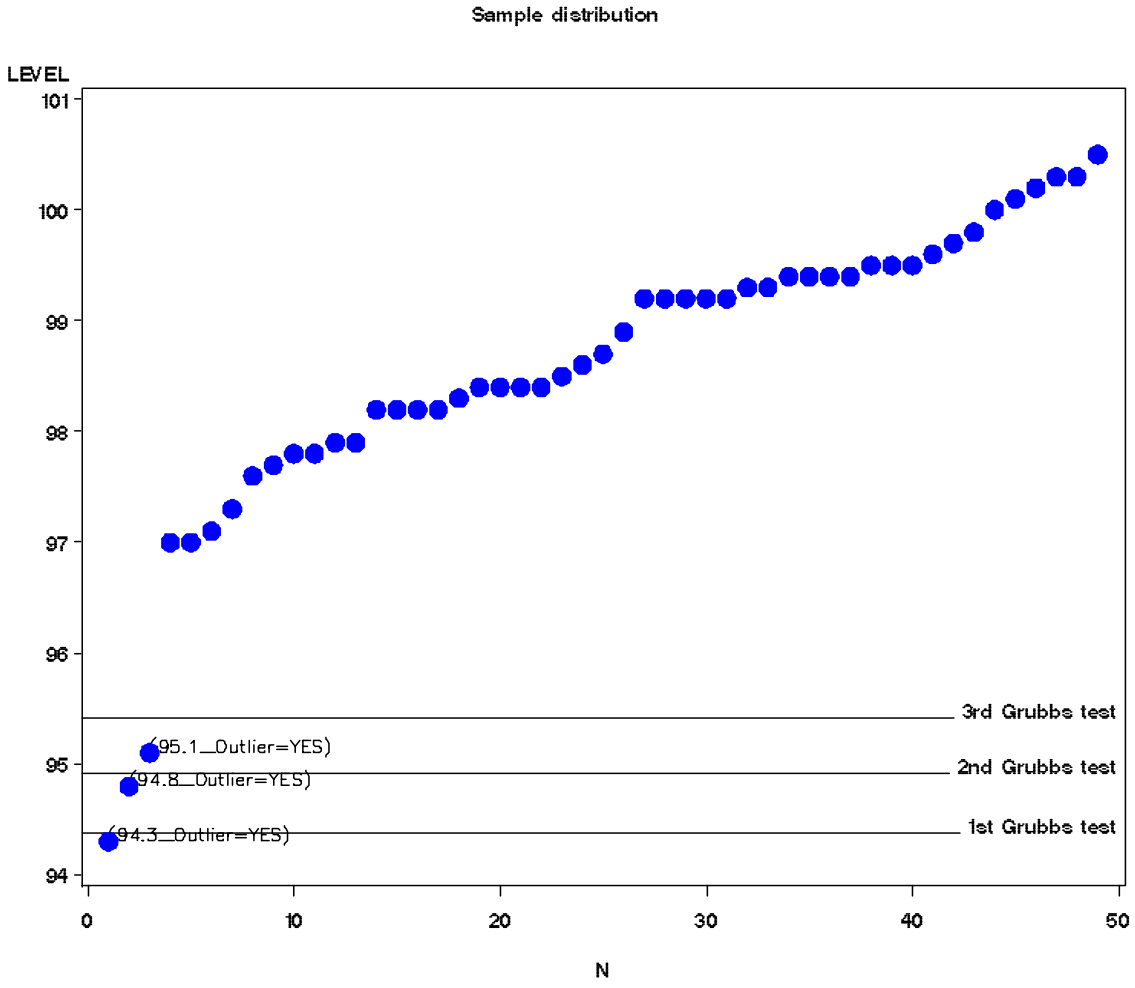


Standarized PROC ROBUSTREG residuals (cutoff = 3.0000)			
Method = M	Method = LTS	Method = S	Method = MM
-3.6745	-3.6377	-3.5731	-3.5549
-3.4895	-3.4562	-3.3969	-3.3790
-3.2119	-3.1839	-3.1327	-3.1150

More detailed results are provided below using PROC ROBUSTREG (Method = M).

Level	Robustreg Standarized Residual	Robustreg Outlier	Sample Mean	Sample Std	Grubbs Critical Value	Grubbs Outlier	Grubbs Statistics
94.8	-3.674515	1	98.587755	1.2909352	3.1201277	0	2.9341172
95	-3.489451	1	98.666667	1.179133	3.1117965	0	3.1096294
95.3	-3.211855	1	98.744681	1.0592878	3.1032431	1	3.2518838

Again, slight, modification of the first measurement in the sample (94.3, 94.8, 95.1) results in successful Grubbs outlier detection.



Standarized PROC ROBUSTREG residuals (cutoff = 3.0000)			
Method = M	Method = LTS	Method = S	Method = MM
-4.3681	-4.0915	-4.0135	-3.9974
-3.8832	-3.6377	-3.5731	-3.5575
-3.5923	-3.3654	-3.3088	-3.2936

More detailed results are provided below using PROC ROBUSTREG (Method = M).

Level	Robustreg Standarized Residual	Robustreg Outlier	Sample Mean	Sample Std	Grubbs Critical Value	Grubbs Outlier	Grubbs Statistics
94.3	-4.368065	1	98.569388	1.3450843	3.1201277	1	3.1740671
94.8	-3.883204	1	98.658333	1.2049249	3.1117965	1	3.202136
95.1	-3.592288	1	98.740426	1.0737297	3.1032431	1	3.3904488

The example above shows, that sequential application of Grubbs test to evaluate multiple outliers may not always correctly identify them due to masking phenomenon. In particular, if the first observation (out of a few lumped, suspicious measurements) is not confirmed as an outlier by Grubbs's test, application of PROC ROBUSTREG may be useful to confirm the finding. (Beckman and Cook 1983, p. 130) noted that power of consecutive applications of Grubbs's test for outliers' decreases as the two outliers move further from the mean. In case of a single outlier, (Beckman and Cook 1983, p. 129), Grubbs's test maximizes the probability of making a correct decision under the mean-shift model.

Please note that the following code SAS® has been applied to plot sample distribution:

```
data labels;
  length function style $8 text $20;
  retain function 'label' xsys ysys '2' hsys '1' style 'arial';
  drop level;

    set fin_out end=lastob;
    x=N; y=level;
    text = compress('('||level||"_Outlier = "||put(outlier,outlier.)||')');
    position = 'C';

    if _N_ = 1 then do;
    position = 'A';
    text = compress('('||"Outlier = "||put(outlier, outlier.)||')');
                  end;

   if outlier >.;
run;

proc gplot data=fin_out;
     plot level*N/overlay anno=labels
     hminor = 0
      vminor = 0
      vaxis = axis1
      vref = (&ref1, &ref2, &ref3)
      frame;
run;
quit;
```

WEISBERG AND 3σ TEST VERSUS ROBUSTREG PROCEDURE

Weisberg t-test to detect outliers was introduced in (Weisberg 2005, Chapter 9). Assume that y_1, y_2 , …, y_n are random observations from a normal distribution **Y**. Weisberg t-test can be performed by fitting a linear regression model with **Y** as the outcome variable and a binary variable **U** is coded as 1 for the suspected observation and 0 otherwise. The p-value for testing whether the regression coefficient **β** for binary variable **U** is 0 will be compared with the adjusted α level (α/n), and if the p-value is equal to or smaller than α/n, then the suspected outlier is confirmed as an outlier. It has been proven in (Pan Dey Bowers and Solak 2008) that Weisberg t-test and Grubbs test are both identical.

Another popular outlier test is 3σ test. This test is based on a fundamental principle of a normal distribution N(μ, σ): 99.7 % of the area under normal curve is included within interval [μ - 3σ, μ + 3σ]. On the other hand, according to (1), 3σ test may be viewed as a special case of Grubbs test with a single critical value of 3. A summary of application 3σ method versus ROBUSTREG procedure is summarized below.

Robustreg procedure using Method=M vs. 3σ test

Level	Robustreg Standarized Residual	Robustreg Outlier	Sample Mean	Sample Std	3sigma Critical Value	3sigma Outlier	3sigma Statistics
94.8	-3.773477	1	98.579592	1.3135479	3	0	2.8773916
94.8	-3.773477	1	98.658333	1.2049249	3	1	3.202136
95.1	-3.489564	1	98.740426	1.0737297	3	1	3.3904488
94.8	-3.674515	1	98.587755	1.2909352	3	0	2.9341172
95	-3.489451	1	98.666667	1.179133	3	1	3.1096294
95.3	-3.211855	1	98.744681	1.0592878	3	1	3.2518838
94.3	-4.368065	1	98.569388	1.3450843	3	1	3.1740671
94.8	-3.883204	1	98.658333	1.2049249	3	1	3.202136
95.1	-3.592288	1	98.740426	1.0737297	3	1	3.3904488

It is interesting to notice that 3σ test, applied sequentially to (94.8, 95.0, 95.3) identified 95.0 as an outlier contrary to Grubbs test. 3σ test can be viewed as a special case of Grubbs test with constant critical value (=3) independent on sample size. In case of sample size N=37 Grubbs critical value best matches 3. For sample sizes <37, 3 is an upper bound on critical values while for sample sizes >37, 3 is a lower bound. Therefore, in the case of 95.0, an outlier confirmed by 3σ test, may not be a Grubbs outlier.

CONCLUSION

The paper compared Grubbs test sequential application to detect outliers in univariate data sets with results provided by SAS®/STAT ROBUSTREG procedure. Multiple outliers, close to each other and away from the main body of data, may not be detected correctly by Grubbs test itself, due to masking phenomenon. It appears that PROC ROBUSTREG may be a useful tool to confirm Grubbs test findings or to detect outliers, missed by it. Grubbs test is a very useful tool in handling single outliers, but has to be applied with care, when multiple outliers are present. Grubbs provided a statistics to handle two outliers at the same time, but for limited sample sizes (4-20) only.

REFERENCES

Beckman, R. J., Cook R. D., 1983, *Outlier…….s*, Technometrics, vol. 25, No. 2 , May, pp. 119-149.

Chen C., 2002, *Robust Regression and Outlier Detection with the ROBUSTREG Procedure*, SUGI 27 Proceedings, Orlando, FL, April 14-17, paper 265-27.

Dixon W. J., 1953, *Processing Data for Outliers,* Biometrics, vol. 9, pp. 74-89

Grubbs F. E., 1950, *Sample Criteria for Testing Outlying Observations*, Annals of Math. Statistics, vol. 21, pp. 27-58.

Grubbs F. E., 1969, *Procedures for Detecting Outlying Observations in Samples*, Technometrics, vol. 11, No. 1, pp. 13-14.

Pan Z., Dey M., Bowers J. S., Solak M. K., 2008, *Tests for Outliers*, Schering-Plough Research Institute, Pharmaceutical Sciences Technical Report.

Weisberg S., 2005*, Applied Linear Regression*, Willey Interscience, 3rd edition.

CONTACT INFORMATION

Your comments and questions are encouraged. Contact the author:

Marek K. Solak, PhD
Schering-Plough Research Institute
556 Morris Ave.
Summit, NJ 07901, Blgd. S-7 A2-2145
Tel.: (908) – 473 – 2884
Fax: (908) – 473 – 5850
E-mail: mark.solak@spcorp.com

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

Other brand and product names are trademarks of respective companies.