

A Technique to Rescue Non-Parametric Outlier Data Using SAS®

Venu Perla Ph.D., Independent SAS Programmer, Cross Lanes, WV 25313

Abstract

Recently, I have published a paper, 'How PROC SQL and SAS® macro programming made my statistical analysis easy? A case study on linear regression' (*Refer CinSUG website for Ohio SAS Users Conference 2015 for a full paper*). In that paper, various macro programs were created to eliminate outlier data during normalization of non-parametric data. Often, this outlier data is valuable and provides a different outlook while drawing conclusions from the entire set of data after analysis. The objective of this paper is to show a technique of rescuing non-parametric outlier data using a non-parametric test after the analysis of parametric portion of the data using SAS. This paper also explains how conclusions are drawn from both, parametric and non-parametric tests without sacrificing the outlier data.

1. Introduction

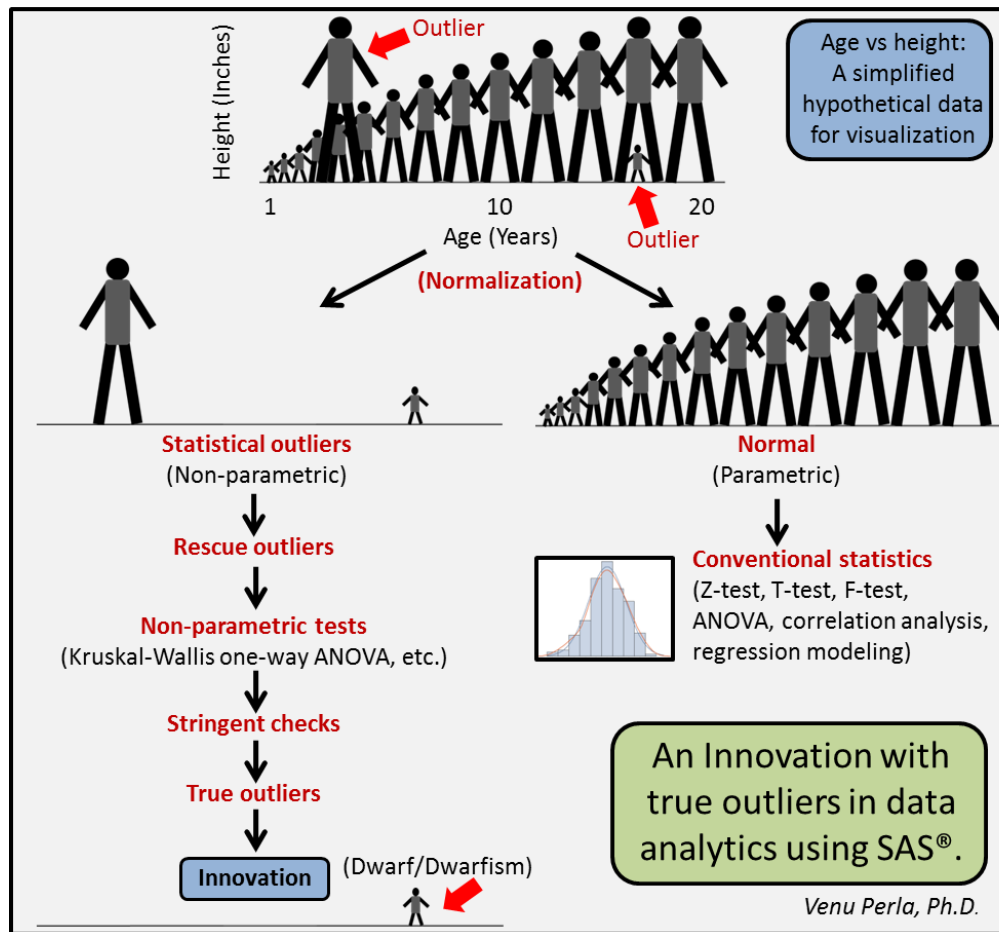
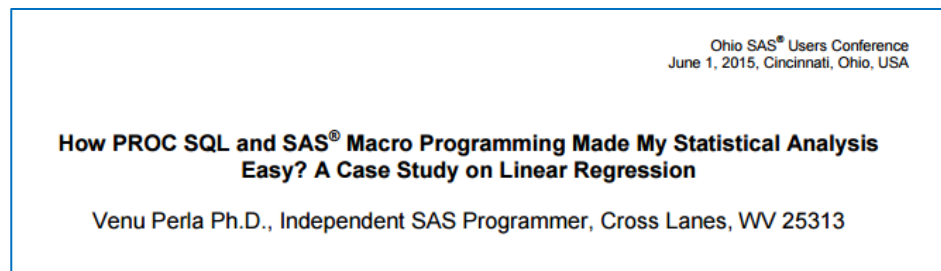


Figure 1

For many scientists and data analysts, outliers are like a 'black box' in conventional statistics. Many believe that these outlier observations arise due to errors or due to improper procedures in the experiment. Majority of them eliminate the outliers unscientifically by brute force. Some identify them statistically but discard them as if they are junk. Some understand importance of the outliers but they do not know how to deal with it. If you are one among them or interested in scope of the outliers, then this paper is the right resource for you. Outliers are like hidden treasure in data analytics. Discarding true outliers from data may costs huge amount of money in certain projects such as clinical

trials. An innovation with true outliers in data analytics using SAS is shown in **Figure 1**. In this figure, a simplified hypothetical data on age vs height is shown for visualization. Two statistical outliers are clearly marked in the pictograph. At about 3 years of age, one outlier's height is equal to 18-20 years old person. As opposed to this, at about 18 years of age, another outlier's height is approximately equal to 3 years old kid. This dataset can be normalized by eliminating these two statistical outliers. As usual, conventional statistics can be performed on the normalized dataset. On the other hand, rescued statistical outliers can be tested with non-parametric tests such as Kruskal-Wallis one-way ANOVA. Stringent checks can be performed on outcome and true outliers can be identified. For example, in this hypothetical dataset, out of two, one outlier can be a dwarf. In medical field, this condition is called as dwarfism. In other words, an 18 years old person with height equal to a 3 years old kid is not an experimental or procedural error. This is a simple hypothetical example of an innovation with true outliers. On the other hand, it is possible that the 3 years old kid with height equal to 18 years appears to be an experimental or procedural error or an error during data processing.

In this paper, a raw data on two interrelated plant metabolites (X and Y) is utilized for analysis. This data is obtained from SHEET1 of DATAXY.XLS, a Microsoft Excel 97-2003 file (See **Table 1** and **Excel SHEET1 snapshot** in *Appendix*). This raw data is transformed, normalized and utilized for generating a linear regression model. While normalization, outlier data is separated and rescued for further non-parametric analysis using Kruskal-Wallis test. Finally, conclusions are drawn from both, parametric and non-parametric portions of the data. Significance and applications of rescuing outlier non-parametric data are discussed. Analysis was carried out using SAS[®] 9.4 software in Windows operating system. To reduce coding, various SAS macros created earlier were utilized in this paper (Perla, 2005). Refer following paper for background and SAS macros.



Terminology Used In This Paper

Outlier

According to the Wikipedia, in statistics, an outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.

Statistical outlier

Since 'Outlier' is a generic term, I prefer to use 'Statistical outlier' for any outlier that is identified with statistical methods.

True outlier

In my opinion, any statistical outlier that is not aroused due to experimental or procedural errors is called as true outlier.

2. Data Import

Macro variable 'PATH' is created for Excel file path. Macro 'EXCEL_IMPORT' utilizes macro variable 'PATH' while importing Excel file 'DATAXY' into SAS. The output SAS dataset created after execution of macro 'EXCEL_IMPORT' is 'XY' (**Table 2**). All the defined macros are stored under the folder 'STATMACROS'. These macros are invoked by 'OPTIONS' statement mentioned below.

```
%let path= C:\Users\Documents\SAS\My SAS Files\;
options mautosource sasautos="C:\Users\Documents\SAS\statmacros\";

%excel_import (excel_file=dataxy, excel_sheet=sheet1, dataset=xy);
**Refer Perla (2015) for macro code;
```

Table 2

Dataset from imported excel data

Obs	ID	X	Y
1	1	0.16	0.16
2	1	0.1444	0.1681
3	1	0.1681	0.1681
4	1	0.1521	0.1521
5	2	0.25	0.25

3. Preliminary Data Analysis

Execution of macro 'SCATTER_CORR' produces scatter plot and Pearson correlation coefficients for X- and Y-variables of the dataset 'XY'. Output indicates that the relationship between X and Y is weak (Pearson correlation coefficient: 0.15) (Figure 2; Table 3).

```
%scatter_corr (dataset=xy, xvar=x, yvar=y);
**Refer Perla (2015) for macro code;
```

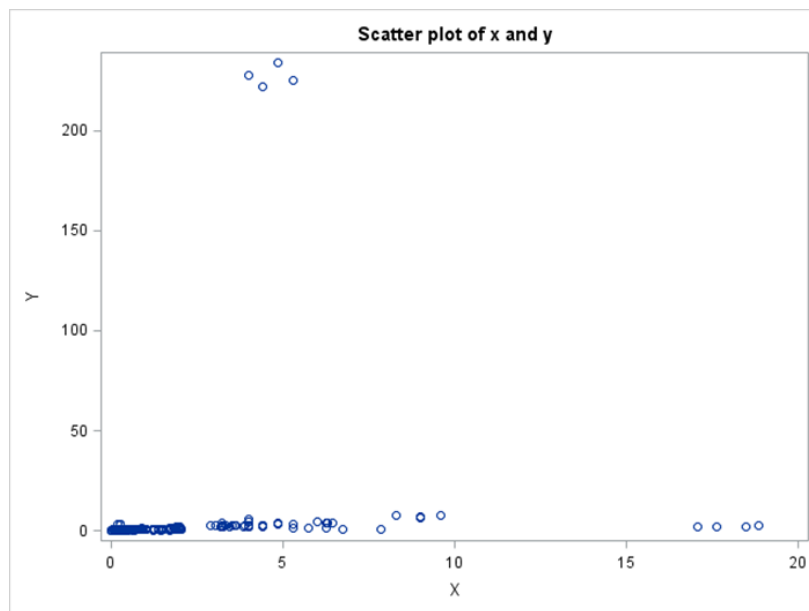


Figure 2

Table 3

Pearson Correlation Coefficients, N = 204 Prob > r under H0: Rho=0		
	X	Y
X	1.00000	0.14778
Y	0.14778	1.00000
	Prob > r	Prob > r
X		0.0349
Y	0.0349	

Output produced after execution of the macro 'REG_NORMALITY' suggests that all the tests of normality are significant and the distribution of residuals for Y is not normal (Figure 3; Table 6). Although, the adjusted R² value is negligible, 'LACK OF FIT' for linear model is not significant. This indicates that the relationship between X and Y can

be explained by linear regression model (**Table 4** and **5**). However, data has to be normalized before developing a linear regression model for X and Y in conventional statistics.

```
%reg_normality (dataset=xy, xvar=x, yvar=y);
**Refer Perla (2015) for macro code;
```

Table 4

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4387.37258	4387.37258	4.51	0.0349
Error	202	196508	972.81120		
Lack of Fit	100	46031	460.30697	0.31	1.0000
Pure Error	102	150477	1475.26633		
Corrected Total	203	200895			

Table 5A

Root MSE	31.18992	R-Square	0.0218
Dependent Mean	5.62215	Adj R-Sq	0.0170
Coeff Var	554.76867		

Table 5B

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	2.38427	2.66332	0.90	0.3717
X	X	1	1.56696	0.73785	2.12	0.0349

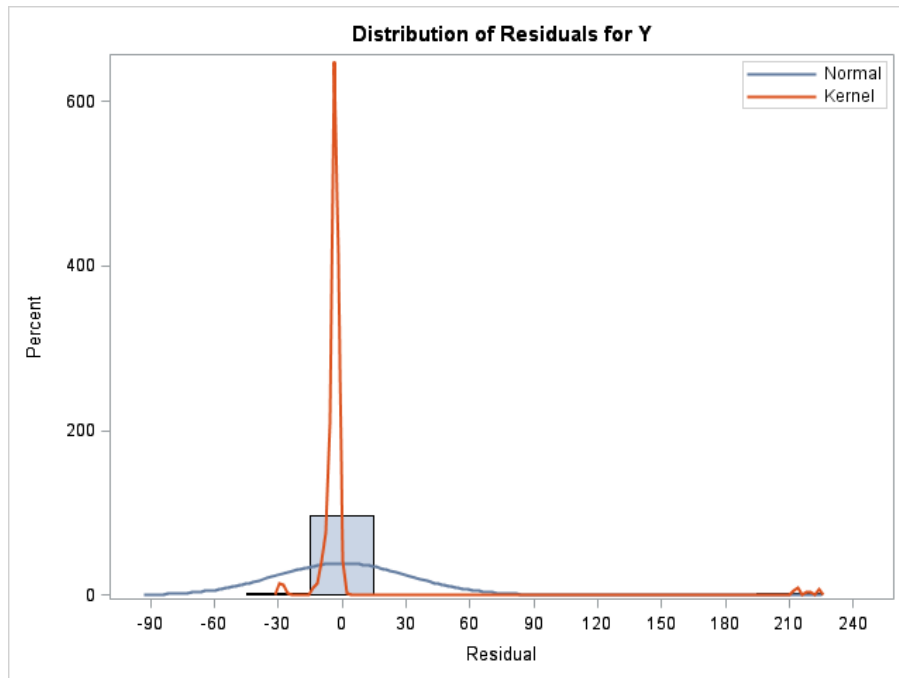


Figure 3

Table 6

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.185463	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.486375	Pr > D	<0.0100
Cramer-von Mises	W-Sq	13.4026	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	64.21648	Pr > A-Sq	<0.0050

4. Data Normalization

4.1. Data Transformation

Data transformation was carried out using Box-Cox power transformation technique (Perla, 2015). The best lambda, which is -0.25, is identified by invoking the macro 'BOX COX LAMBDA' on non-zero and non-negative dataset 'XY' (Table 7). This lambda is the exponent to be used for transforming the Y-variable.

```
%box_cox_lambda (pre_trans_dataset=xy, xvar=x, yvar=y);
**Refer Perla (2015) for macro code;
```

Table 7

Box-Cox Transformation Information for Y			
Lambda	R-Square	Log Like	
-3.00	0.05	-457.67	
-2.75	0.06	-388.06	
-2.50	0.07	-320.00	
-0.25	0.33	72.15	<

< - Best Lambda
 * - 95% Confidence Interval
 + - Convenient Lambda

In biological sciences, it is advisable to adopt a transformation that is commonly used by others in the field. Since the best lambda value is close to '0', logarithmic transformation is adopted for Y-variable in this paper. Macro 'TRANSFORM_LAMBDA_2,' a modified version of macro 'TRANSFORM_LAMBDA' (Perla, 2015), is explained in *Appendix*. Global macro variable 'OTHERVARS' is a comma separated list of variables to be included in the output dataset 'XY_TRANS'. Variable 'ZERO_Y' in the output dataset represents Y-variable after logarithmic transformation (Table 8). Variable 'ZERO_Y' is used for further analysis.

```
%let othervars=id,x;
%transform_lambda_2 (pre_trans_dataset=xy, yvar=y, trans_dataset= xy_trans);
```

Table 8

Transformation of y-variable with convenient lambda										
Obs	ID	X	Y	neg_2_y	neg_1_y	neg_half_y	zero_y	half_y	one_y	two_y
1	1	0.16	0.16	39.0625	6.25000	2.50000	-1.83258	0.40	0.160	0.03
2	1	0.1444	0.1681	35.3887	5.94884	2.43902	-1.78320	0.41	0.168	0.03
3	1	0.1681	0.1681	35.3887	5.94884	2.43902	-1.78320	0.41	0.168	0.03

For a meaningful Y-intercept, X-variable is standardized using the macro 'STDIZE_X'. The dataset 'XY_TRANS' obtained above is used here to get output dataset 'XY1'. 'XY1' contains transformed Y-variable and standardized X-variable (X) (Table 9).

```
%stdize_x (trans_dataset=xy_trans, trans_stdize_dataset=xy1, xvar=x);
**Refer Perla (2015) for macro code;
```

Table 9

Standardized x-variable after Y-transformation											
Obs	ID	Unstdized_X	Y	neg_2_y	neg_1_y	neg_half_y	zero_y	half_y	one_y	two_y	X
1	1	0.16	0.16	39.0625	6.25000	2.50000	-1.83258	0.40	0.160	0.03	-1.9063461
2	1	0.1444	0.1681	35.3887	5.94884	2.43902	-1.78320	0.41	0.168	0.03	-1.9219461
3	1	0.1681	0.1681	35.3887	5.94884	2.43902	-1.78320	0.41	0.168	0.03	-1.9063461

4.2. Identification and Slicing of Statistical Outliers While Normalizing the Data

Macro 'REGRESSION_WOUT_OUTLIERS_2,' a modified version of macro 'REGRESSION_WOUT_OUTLIERS' (Perla, 2015), is explained in *Appendix*. This master macro generates ANOVA, parametric estimates, distribution of residuals and tests of normality for the input dataset (DATASET or INDATA). Furthermore, output generated on outlier and influential observations is useful while identifying and slicing of outlier observations from the input dataset. There are four macros (REG_NORMALITY, OUTLIER_OBS, SLICE_OBS_2 and NO_OUTLIER_DATA) within this master macro. Except for macro 'SLICE_OBS_2', all the keyword parameters for other macros are described earlier by Perla (2015). Macro 'SLICE_OBS_2' is explained in *Appendix*. 'R, INFLUENCE, RSTUDENTBYLEVERAGE, DFFITS, DFBETAS' and 'COOKSD' options of the macro 'OUTLIER_OBS' generate detailed outlier and or influential observations for the input dataset. Outliers are sliced as per the criteria adopted by Perla (2015).

Set of outlier observations to be sliced are referenced by a global macro variable 'OBSET'. Master macro 'REGRESSION_WOUT_OUTLIERS_2' is invoked twice, first without outliers (OBSET=0), then with the identified outliers from the first run. Outlier free dataset is generated after second run. This two-step process is continued until all the tests of normality and 'LACK OF FIT' options produce non-significant values. Perhaps, one should be careful while identifying and slicing statistical outliers in the dataset.

```
%let obset=0;
%regression_wout_outliers_2 (dataset=xy1, indata=xy1, sliced_data=sliced,
outdata=xy2, xvar=x, yvar=zero_y);
**From this run, it is clear that obs=9,10,11,12,169,170,171,172 are outliers;
%let obset=9,10,11,12,169,170,171,172;
**For improvement, slice obs=9,10,11,12,169,170,171,172 from XY1 dataset to get
outlier free data;
%regression_wout_outliers_2 (dataset=xy1, indata=xy1, sliced_data=sliced,
outdata=xy2, xvar=x, yvar=zero_y);
```

In the first cycle of execution shown above, results generated with 'OBSET=0' indicates that the transformed dataset is not normal (**Figure 4; Table 10**). However, model still holds good for X- and Y-variables (**Table 11**). As dataset is not normal, observation numbers for statistical outliers are identified from the results (**Table 12; Figures 5-8**), and the master macro is executed with 'OBSET=9, 10, 11, 12, 169, 170, 171, 172'. Outlier-free output dataset (XY2) is tested again in the next cycle of execution of master macro. After several cycles of execution, normality tests and 'LACK OF FIT' values for dataset 'XY5' are non-significant (**Table 13 and 14; Figure 9**). Either dataset 'XY5' or 'XY6' can be used for conventional statistics.

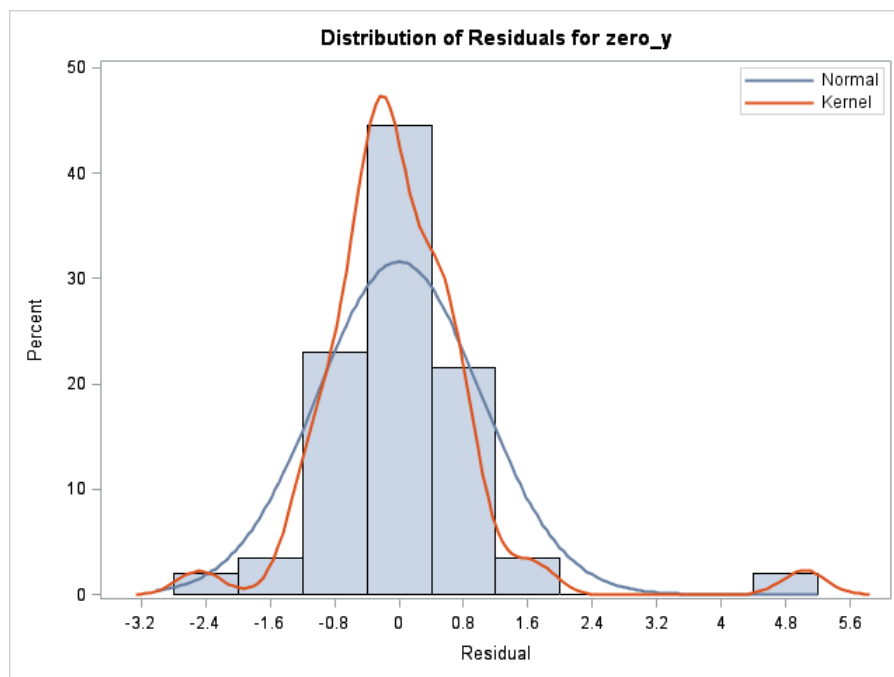


Figure 4

Table 10

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.807399	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.119701	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.021952	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	6.879378	Pr > A-Sq	<0.0050

Table 11

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	84.10059	84.10059	81.99	<.0001
Error	202	207.18810	1.02568		
Lack of Fit	100	118.42684	1.18427	1.36	0.0614
Pure Error	102	88.76126	0.87021		
Corrected Total	203	291.28870			

Table 12

The REG Procedure
 Model: MODEL1
 Dependent Variable: zero_y

Output Statistics																	
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2 -1 0 1 2				Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITs	DFBETAS	
																Intercept	X
9	5.4557	0.4407	0.0972	5.0150	1.008	4.975					0.115	5.2975	0.0092	0.7849	0.5107	0.3726	0.3492
10	5.4027	0.3474	0.0904	5.0553	1.009	5.012					0.101	5.3423	0.0080	0.7807	0.4790	0.3755	0.2974
11	5.4294	0.2585	0.0847	5.1709	1.009	5.124					0.092	5.4794	0.0070	0.7699	0.4599	0.3850	0.2515
12	5.4161	0.5384	0.1048	4.8777	1.007	4.842					0.127	5.1377	0.0107	0.7977	0.5348	0.3617	0.3939
169	0.8109	3.2177	0.3798	-2.4067	0.939	-2.564					0.538	-2.5998	0.1406	1.1000	-1.0517	-0.1964	-1.0332
170	0.8242	3.4021	0.3998	-2.5779	0.930	-2.770					0.709	-2.8176	0.1559	1.1073	-1.2107	-0.2147	-1.1915
171	0.7705	3.0912	0.3661	-2.3206	0.944	-2.458					0.454	-2.4890	0.1307	1.0933	-0.9649	-0.1869	-0.9467
172	0.8765	3.4770	0.4080	-2.6005	0.927	-2.805					0.762	-2.8547	0.1623	1.1135	-1.2564	-0.2184	-1.2373

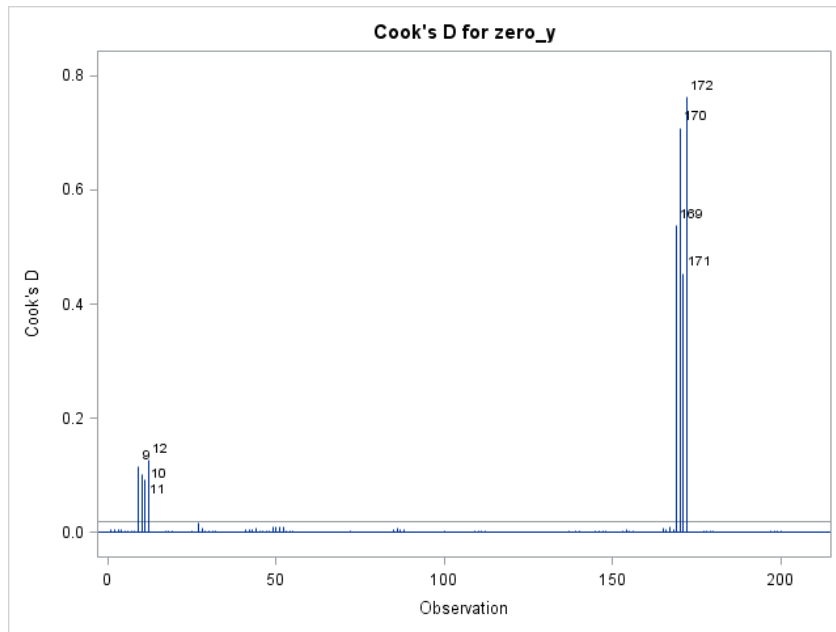


Figure 5

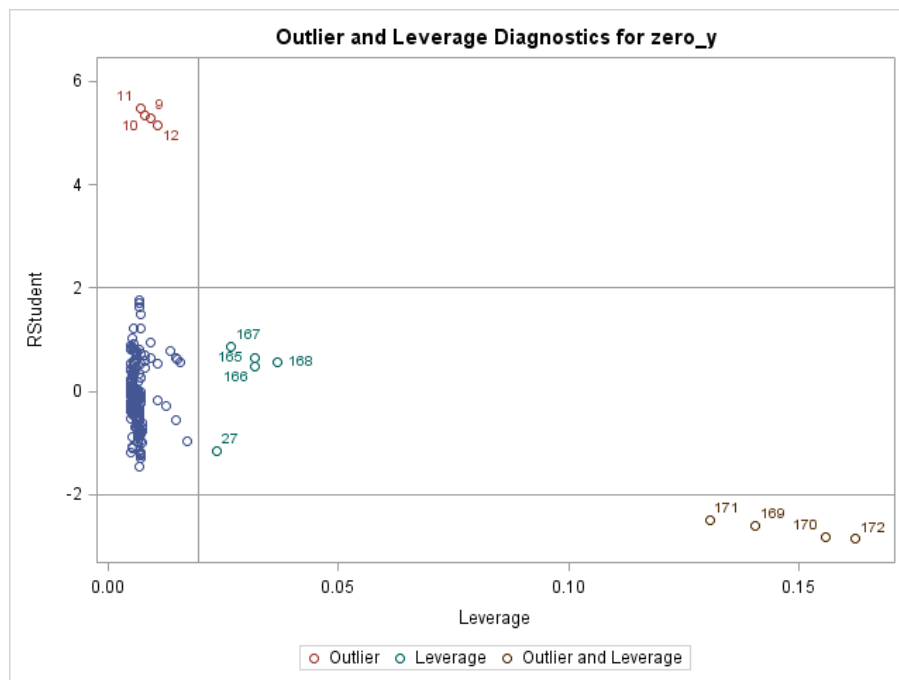


Figure 6

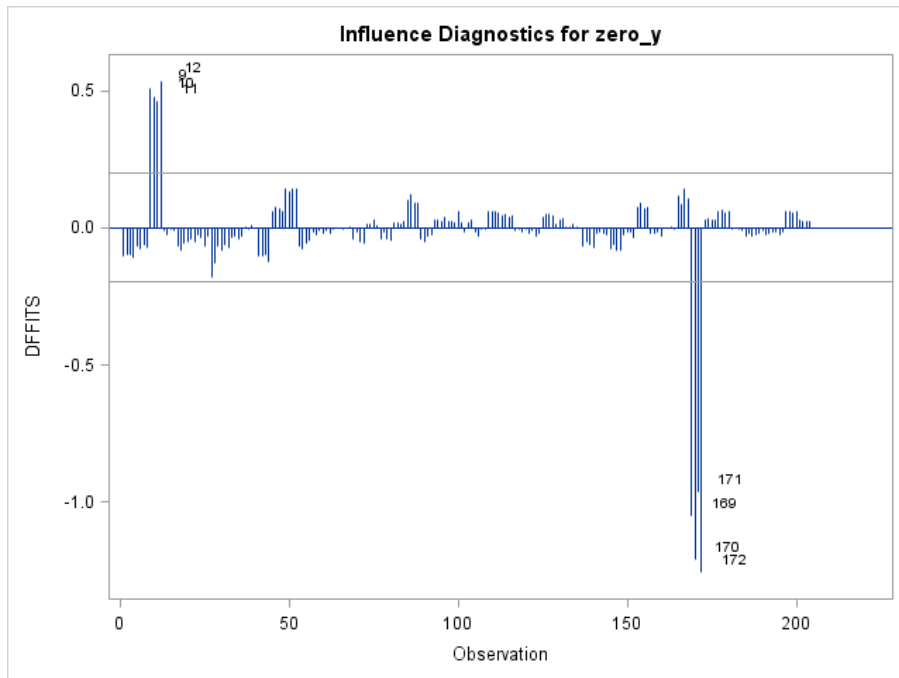


Figure 7

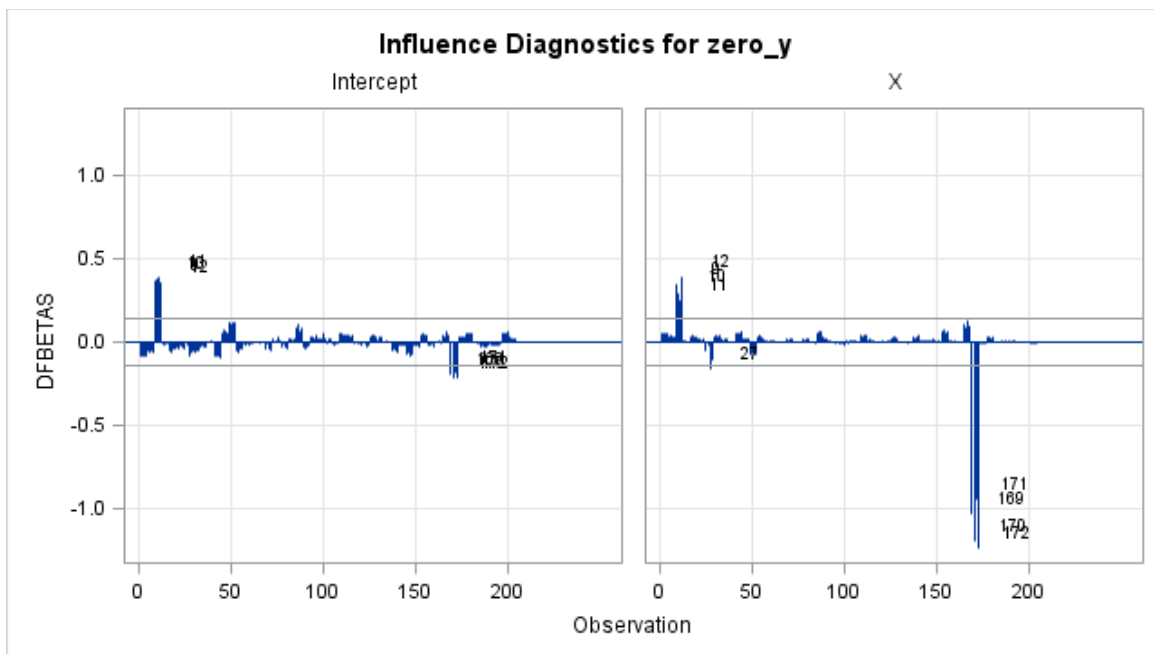


Figure 8

```
%let obset=0;
%regression_wout_outliers_2 (dataset=xy2, indata=xy2, sliced_data=sliced1,
outdata=xy3, xvar=x, yvar=zero_y);
**From this run, it is clear that obs=21,22,23,24 are outliers;
%let obset=21,22,23,24;
**For improvement, slice obs=21,22,23,24 from XY2 dataset to get outlier free
data;
%regression_wout_outliers_2 (dataset=xy2, indata=xy2, sliced_data=sliced1,
outdata=xy3, xvar=x, yvar=zero_y);
```

```

%let obset=0;
%regression_wout_outliers_2 (dataset=xy3, indata=xy3, sliced_data=sliced2,
outdata=xy4, xvar=x, yvar=zero_y);
**From this run, it is clear that obs=41,42,43,44 are outliers;
%let obset=41,42,43,44;
**For improvement, slice obs=41,42,43,44 from XY3 dataset to get outlier free
data;
%regression_wout_outliers_2 (dataset=xy3, indata=xy3, sliced_data=sliced2,
outdata=xy4, xvar=x, yvar=zero_y);

%let obset=0;
%regression_wout_outliers_2 (dataset=xy4, indata=xy4, sliced_data=sliced3,
outdata=xy5, xvar=x, yvar=zero_y);
**From this run, it is clear that obs=153,154,155,156 are outliers;
%let obset=153,154,155,156;
**For improvement, slice obs=153,154,155,156 from XY4 dataset to get outlier
free data;
%regression_wout_outliers_2 (dataset=xy4, indata=xy4, sliced_data=sliced3,
outdata=xy5, xvar=x, yvar=zero_y);

%let obset=0;
%regression_wout_outliers_2 (dataset=xy5, indata=xy5, sliced_data=sliced4,
outdata=xy6, xvar=x, yvar=zero_y);
    
```

Table 13

Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	1	93.65421	93.65421	499.56	<.0001	
Error	182	34.12002	0.18747			
Lack of Fit	90	18.16991	0.20189	1.16	0.2343	
Pure Error	92	15.95011	0.17337			
Corrected Total	183	127.77423				

Root MSE	0.43298	R-Square	0.7330
Dependent Mean	-0.38424	Adj R-Sq	0.7315
Coeff Var	-112.68624		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	-0.08040	0.03469	-2.32	0.0216
X	X	1	0.49608	0.02220	22.35	<.0001

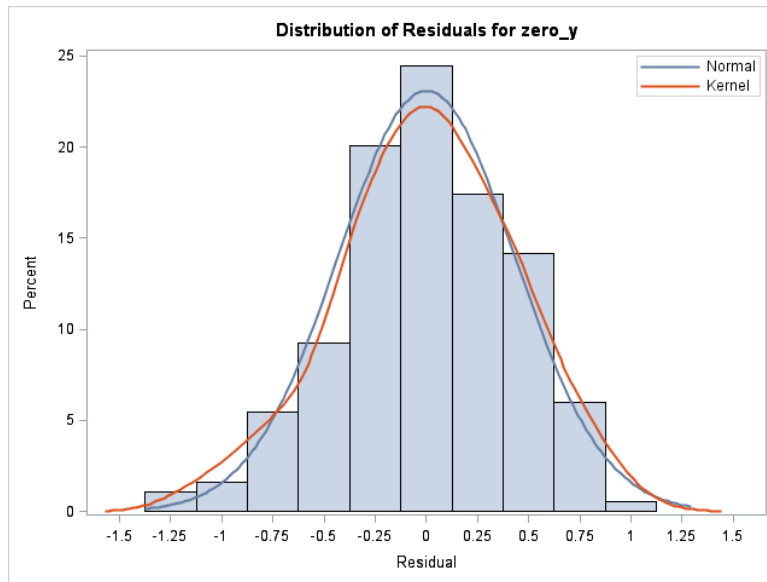


Figure 9

Table 14

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.992525	Pr < W	0.4666
Kolmogorov-Smirnov	D	0.032933	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.02707	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.250559	Pr > A-Sq	>0.2500

5. Linear Regression Model

'PARAMETER ESTIMATES' obtained from dataset 'XY5' after execution of the macro 'REGRESSION_WOUT_OUTLIERS_2' are used for modelling linear regression. Note that adjusted R^2 value is increased from 0.02 to 0.73 after normalization of data by slicing 5 IDs (5 IDs X 4 Replications = 20 Observations). The relationship between X- and Y-variables in the parametric portion of data can be explained by linear regression model (2).

1. Raw data (non-normal): $Y = 2.384 + 1.567X$ (Adjusted R^2 : 0.02)
2. Normalized data (without 20 observations): $Y = -0.080 + 0.496X$ (Adjusted R^2 : 0.73)

6. Correlation Analysis of Parametric Portion of the Data

Correlation analysis of normalized portion of the data is performed by calling again the macro 'SCATTER CORR' for UNSTDIZED_X- and Y-variables of the dataset 'XY6'. Note that Pearson correlation coefficient of raw data (Figure 2; Table 3) is increased from 0.15 to 0.88 after normalization (Figure 10; Table 15).

```
%scatter_corr (dataset=xy6, xvar=unstdized_x, yvar=y);
**Refer Perla (2015) for macro code;
```

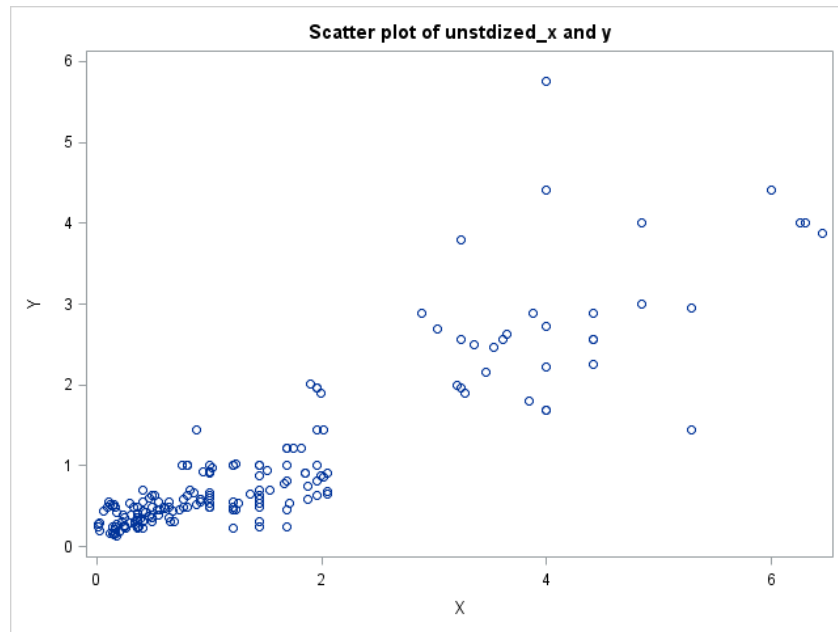


Figure 10

Table 15

Pearson Correlation Coefficients, N = 184 Prob > r under H0: Rho=0		
	Unstdized_X	Y
Unstdized_X	1.00000	0.88337
X		<.0001
Y	0.88337	1.00000
Y	<.0001	

7. Rescuing and Analyzing Statistical Outliers

There are 204 observations in the raw data set. Out of which 20 are outliers. From the above normalization process, it is clear that approximately 10% of the observations are outliers in the raw data. This outlier data is generally ignored when predictions are made for whole data. In depth analysis of outliers may provide meaningful insights on the data. Often, all the outliers are not true outliers. For this reason, the sliced 20 outliers (5 IDs) are combined, sorted and analyzed using non-parametric one-way ANOVA (Kruskall-Wallis test) after incorporating one maximum and one minimum ID from parametric portion of the data for reference. Details of this analysis are given below separately for X- and Y-variable.

7.1. Combining All the Outlier Datasets

Outlier datasets (SLICED-SLICED3) are combined and sorted to get 'SLICED_ALL' dataset (Table 16). There are 20 observations (5 IDs) in the 'SLICED_ALL' dataset.

```

title "Combined dataset for all the outliers from SLICED-SLICED4 datasets";
data sliced_all;
    set sliced sliced1 sliced2 sliced3;
run;

Proc sort data=sliced_all;
    by id;
run;

proc print data=sliced_all;
run;
    
```

Table 16

Obs	ID	Unstdized_X	Y	neg_2_y	neg_1_y	neg_half_y	zero_y	half_y	one_y	two_y	X
1	3	4.84	234.09	0.00002	0.00427	0.06536	5.45571	15.30	234.090	54798.13	2.77365392
2	3	4.41	222.01	0.00002	0.00450	0.06711	5.40272	14.90	222.010	49288.44	2.34365392
3	3	4	228.01	0.00002	0.00439	0.06623	5.42939	15.10	228.010	51988.56	1.93365392
4	3	5.29	225	0.00002	0.00444	0.06667	5.41610	15.00	225.000	50625.00	3.22365392
5	7	5.76	1.4161	0.49867	0.70616	0.84034	0.34791	1.19	1.416	2.01	3.69365392
6	7	6.25	1.21	0.68301	0.82645	0.90909	0.19062	1.10	1.210	1.46	4.18365392
7	7	6.76	0.8836	1.28082	1.13173	1.06383	-0.12375	0.94	0.884	0.78	4.69365392
8	7	7.84	0.9409	1.12957	1.06281	1.03093	-0.06092	0.97	0.941	0.89	5.77365392
9	13	0.1936	3.1329	0.10188	0.31919	0.56497	1.14196	1.77	3.133	9.82	-1.8727461
10	13	0.25	2.9241	0.11695	0.34199	0.58480	1.07299	1.71	2.924	8.55	-1.8163461
11	13	0.25	3.24	0.09526	0.30864	0.55556	1.17557	1.80	3.240	10.50	-1.8163461
12	13	0.25	3.3489	0.08917	0.29861	0.54645	1.20863	1.83	3.349	11.22	-1.8163461
13	42	8.2944	7.8961	0.01604	0.12664	0.35587	2.06637	2.81	7.896	62.35	6.22805392
14	42	9	6.25	0.02560	0.16000	0.40000	1.83258	2.50	6.250	39.06	6.93365392
15	42	9	7.29	0.01882	0.13717	0.37037	1.98650	2.70	7.290	53.14	6.93365392
16	42	9.61	7.6176	0.01723	0.13127	0.36232	2.03046	2.76	7.618	58.03	7.54365392
17	43	17.64	2.25	0.19753	0.44444	0.66667	0.81093	1.50	2.250	5.06	15.5736539
18	43	18.49	2.2801	0.19235	0.43858	0.66225	0.82422	1.51	2.280	5.20	16.4236539
19	43	17.0569	2.1609	0.21416	0.46277	0.68027	0.77052	1.47	2.161	4.67	14.9905539
20	43	18.8356	2.4025	0.17325	0.41623	0.64516	0.87651	1.55	2.403	5.77	16.7692539

7.2. X-Variable

7.2.1. Identification of Maximum and Minimum IDs for X-Variable in the Parametric Dataset

Maximum and minimum IDs for UNSTDIZED_X-variable in the parametric dataset 'XY6' are identified by using 'MEANS' and 'SORT' procedures of SAS. From the sorted output dataset 'XY6_MAXMIN_X', it is very much evident that ID 39 and ID 5 are the maximum and minimum IDs for UNSTDIZED_X-variable, respectively (Table 17).

```

title "Identification of IDs with max and min X-mean values in parametric
dataset";
proc means data=xy6 mean;
class id;
var unstdized_x;
output out=xy6_maxmin_x mean=x_mean;
run;

Proc sort data=xy6_maxmin_x;
  by descending x_mean;
run;

proc print data=xy6_maxmin_x;
run;

```

Table 17

Obs	ID	_TYPE_	_FREQ_	x_mean
1	39	1	4	6.25105
47	5	1	4	0.01605

7.2.2. Dataset for Two IDs (ID 39 and ID 5)

'PROC SQL' is used to create a table 'X_MAX_MIN' for ID 39 and ID 5 (Table 18).

```

title "Dataset for two IDs with max and min X-values";
proc sql;
create table x_max_min as
select*
from xy6
where id in (39,5);
quit;

proc print data=x_max_min;
run;
    
```

Table 18

Dataset for two IDs with max and min X-values

Obs	ID	Unstdized_X	Y	neg_2_y	neg_1_y	neg_half_y	zero_y	half_y	one_y	two_y	X
1	5	0.01	0.25	16.0000	4.00000	2.00000	-1.38629	0.50	0.2500	0.0625	-2.0563461
2	5	0.0121	0.2704	13.6769	3.69822	1.92308	-1.30785	0.52	0.2704	0.0731	-2.0542461
3	5	0.0196	0.2916	11.7605	3.42936	1.85185	-1.23237	0.54	0.2916	0.0850	-2.0467461
4	5	0.0225	0.2025	24.3865	4.93827	2.22222	-1.59702	0.45	0.2025	0.0410	-2.0438461
5	39	6.0025	4.41	0.0514	0.22676	0.47619	1.48387	2.10	4.4100	19.4481	3.93615392
6	39	6.25	4	0.0625	0.25000	0.50000	1.38629	2.00	4.0000	16.0000	4.18365392
7	39	6.3001	4	0.0625	0.25000	0.50000	1.38629	2.00	4.0000	16.0000	4.23375392
8	39	6.4516	3.8809	0.0664	0.25767	0.50761	1.35607	1.97	3.8809	15.0614	4.38525392

7.2.3. 'X_MAX_MIN' Dataset is Combined with 'SLICED_ALL' Outlier Dataset

Now, two datasets, 'X_MAX_MIN' and 'SLICED_ALL', are merged to get 'SLICED_ALL_2_X_MAX_MIN' dataset. Maximum and minimum X-values of parametric data serve as reference points for statistical outliers in the combined dataset (Table 19). Dataset 'SLICED_ALL_2_X_MAX_MIN' is sorted to maintain the ID numbers in order (Table 20).

```

title "Now add max and min X IDs to dataset SLICED_ALL";
data sliced_all_2_x_max_min;
set sliced_all
x_max_min;
run;
proc print data=sliced_all_2_x_max_min;
run;
    
```

Table 19

Obs	ID	Unstdized_X	Y	neg_2_y	neg_1_y	neg_half_y	zero_y	half_y	one_y	two_y	X
1	3	4.84	234.09	0.0000	0.00427	0.06536	5.45571	15.30	234.090	54798.13	2.77365392
28	39	6.4516	3.8809	0.0664	0.25767	0.50761	1.35607	1.97	3.881	15.06	4.38525392

```

title "To maintain ID order, sort the dataset";
proc sort data=sliced_all_2_x_max_min;
  by id;
run;
proc print data=sliced_all_2_x_max_min;
run;

```

Table 20

Obs	ID	Unstdized_X	Y	neg_2_y	neg_1_y	neg_half_y	zero_y	half_y	one_y	two_y	X
1	3	4.84	234.09	0.0000	0.00427	0.06536	5.45571	15.30	234.090	54798.13	2.77365392
28	43	18.8356	2.4025	0.1732	0.41623	0.64516	0.87651	1.55	2.403	5.77	16.7692539

7.2.4. Kruskal-Wallis Test for Non-Parametric Portion of X-variable

The Kruskal-Wallis one-way ANOVA is performed on 'SLICED_ALL_2_X_MAX_MIN' dataset using 'PROC NPAR1WAY' with 'WILCOXON' option. Distribution of Wilcoxon scores for 'UNSTDIZED_X' is presented in **Figure 11**. There are 7 IDs in the **Figure 11**. Out of which, ID 39 and ID 5 are upper and lower limits from parametric data, respectively. Although, there are 5 outliers (ID 3, ID 7, ID 13, ID 42 and ID 43), only 2 outliers (ID 42 and ID 43) appear to be deviated from normal population of X-variable. Remaining 3 (ID 3, ID 7 and ID 13) are within the range of upper and lower limits of parametric data. In summary, present technique identified only 2 true potential outliers (ID 42 and ID 43) for X-variable.

```

title "Kruskal-Wallis Test for non-parametric portion of X-variable";
title "Two IDs (39 and 5) with maximum and minimum X-values from parametric
portion of the data serve as reference for outliers";
proc npar1way data=sliced_all_2_x_max_min wilcoxon; **non-parametric-one-way;
  class id;
  var unstdized_x;
run;

```

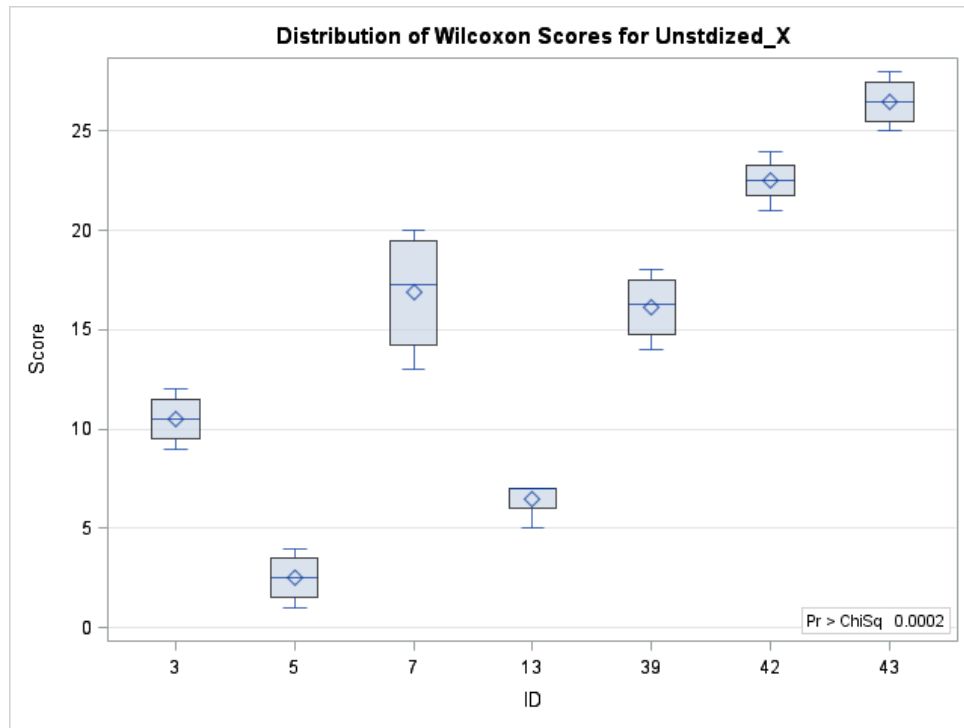



Figure 11

7.3. Y-Variable

Similar to X-variable, analysis for Y-variable is carried out below for statistical outliers.

7.3.1. Identification of Maximum and Minimum IDs for Y-Variable in the Parametric Dataset

The maximum and minimum IDs for Y-variable in the parametric dataset 'XY6' are identified by using 'MEANS' and 'SORT' procedures of SAS. From sorted output dataset 'XY6_MAXMIN_Y', it is very much evident that ID 22 and ID 11 are the maximum and minimum IDs for Y-variable, respectively (Table 21).

```

title "Identification of IDs with max and min Y-mean values in parametric
dataset";
proc means data=xy6 mean;
class id;
var y;
output out=xy6_maxmin_y mean=y_mean;
run;

Proc sort data=xy6_maxmin_y;
by descending y_mean;
run;

proc print data=xy6_maxmin_y;
run;
    
```

Table 21

Obs	ID	_TYPE_	_FREQ_	y_mean
1	22	1	4	4.493125
47	11	1	4	0.1565

7.3.2. Dataset for Two IDs (ID 22 and ID 11)

'PROC SQL' is used to create a table 'Y_MAX_MIN' for ID 22 and ID 11 (Table 22).

```

title "Dataset for two IDs with max and min Y-values";
proc sql;
create table y_max_min as
select*
from xy6
where id in (22,11);
quit;

proc print data=y_max_min;
run;
    
```

Table 22

Dataset for two IDs with max and min Y-values

Obs	ID	Unstdized_X	Y	neg_2_y	neg_1_y	neg_half_y	zero_y	half_y	one_y	two_y	X
1	11	0.1225	0.16	39.0625	6.25000	2.50000	-1.83258	0.40	0.1600	0.0256	-1.9438461
2	11	0.16	0.16	39.0625	6.25000	2.50000	-1.83258	0.40	0.1600	0.0256	-1.9063461
3	11	0.1764	0.1296	59.5374	7.71605	2.77778	-2.04330	0.36	0.1296	0.0168	-1.8899461
4	11	0.1936	0.1764	32.1368	5.66893	2.38095	-1.73500	0.42	0.1764	0.0311	-1.8727461
5	22	3.24	3.8025	0.0692	0.26298	0.51282	1.33566	1.95	3.8025	14.4590	1.17365392
6	22	4	4.41	0.0514	0.22676	0.47619	1.48387	2.10	4.4100	19.4481	1.93365392
7	22	4	5.76	0.0301	0.17361	0.41667	1.75094	2.40	5.7600	33.1776	1.93365392
8	22	4.84	4	0.0625	0.25000	0.50000	1.38629	2.00	4.0000	16.0000	2.77365392

7.3.3. 'Y_MAX_MIN' Dataset is Combined with 'SLICED_ALL' Outlier Dataset

Now, the two datasets, 'Y_MAX_MIN' and 'SLICED_ALL' are combined to get 'SLICED_ALL_2_Y_MAX_MIN' dataset. Maximum and minimum Y-values from parametric data serve as reference points for outliers in the combined dataset (Table 23). Dataset 'SLICED_ALL_2_Y_MAX_MIN' is sorted to maintain the ID numbers in order (Table 24).

```

title "Now add max and min Y IDs to dataset SLICED_ALL";
data sliced_all_2_y_max_min;
set sliced_all
y_max_min;
run;
proc print data=sliced_all_2_y_max_min;
run;
    
```

Table 23

Obs	ID	Unstdized_X	Y	neg_2_y	neg_1_y	neg_half_y	zero_y	half_y	one_y	two_y	X
1	3	4.84	234.09	0.0000	0.00427	0.06536	5.45571	15.30	234.090	54798.13	2.77365392
28	22	4.84	4	0.0625	0.25000	0.50000	1.38629	2.00	4.000	16.00	2.77365392

```

title "To maintain ID order, sort the dataset";
proc sort data=sliced_all_2_y_max_min;
by id;
run;
proc print data=sliced_all_2_y_max_min;
run;
    
```

Table 24

Obs	ID	Unstdized_X	Y	neg_2_y	neg_1_y	neg_half_y	zero_y	half_y	one_y	two_y	X
1	3	4.84	234.09	0.0000	0.00427	0.06536	5.45571	15.30	234.090	54798.13	2.77365392
28	43	18.8356	2.4025	0.1732	0.41623	0.64516	0.87651	1.55	2.403	5.77	16.7692539

7.3.4. Kruskal-Wallis Test for Non-Parametric Portion of Y-variable

The Kruskal-Wallis one-way ANOVA is performed on 'SLICED_ALL_2_Y_MAX_MIN' dataset using 'PROC NPAR1WAY' with 'WILCOXON' option. Distribution of Wilcoxon scores for Y is presented in **Figure 12**. There are 7 IDs in the **Figure 12**. Out of which, ID 22 and ID 11 are upper and lower limits from parametric data, respectively. Although, there are 5 outliers (ID 3, ID 7, ID 13, ID 42 and ID 43), only 2 outliers (ID 3 and ID 42) appear to be deviated from the normal population of Y-variable. Remaining 3 (ID 7, ID13 and ID 43) are within the range of upper and lower limits of parametric data. In summary, present technique identified only 2 potential true outliers (ID 3 and ID 42) for Y-variable.

```

title "Kruskal-Wallis Test for non-parametric portion of Y-variable";
title "Two IDs (22 and 11) with maximum and minimum Y-values from parametric
portion of the data serve as reference for outliers";
proc npar1way data=sliced_all_2_y_max_min wilcoxon; **non-parametric-one-way;
    class id;
    var y;
run;
    
```

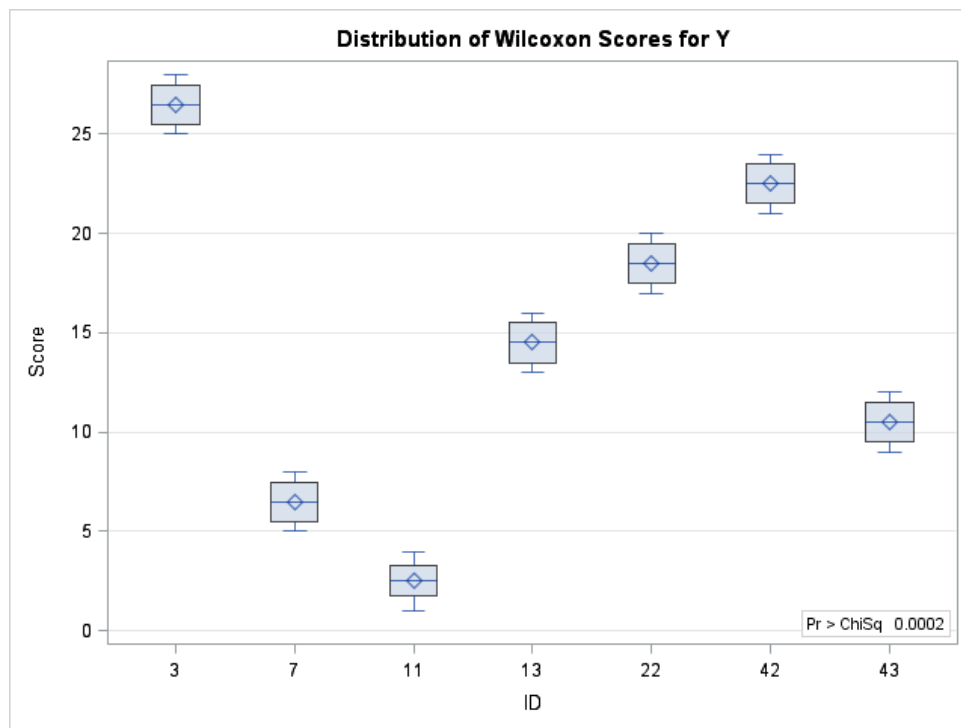


Figure 12

8. Outliers to be Scrutinized

Above outlier analysis for X-variable indicates that ID 42 and ID 43 are two potential outliers to be scrutinized for their X-values in the data. Similarly, outlier analysis for Y-variable indicates that ID 3 and ID 42 are two potential outliers to be scrutinized for their Y-values in the data. Overall, ID 42 appears to be undisputed outlier for both, X- and Y-variables. Following are the partial list of checks to be performed on potential true outliers before confirming them as true outliers:

Errors while entering data.

- Errors while processing the data.
- Errors in wet lab analysis.
- Errors while processing samples for wet lab analysis, and
- Errors in sample collection.

These checks may vary from project to project. If these statistical outliers are not aroused due to errors, then there might be significant reasons for these non-normal values.

9. Applications

In life sciences, true outliers may serve as a source for identifying unique mechanism, pathway, genotype, strain or variety. Similarly, these true outliers may play a role in innovation while analyzing data pertaining to science, clinical research, technology, internet, banking, finance, marketing and other similar sectors.

10. Conclusions

Summary of various SAS programming steps involved in rescuing statistical outliers are presented in **Figure 13**. In this paper, a data on two interrelated plant metabolites was utilized for analysis. Various macros that were previously defined were utilized with or without modifications while importing and normalizing data, and developing a linear regression model for the two plant metabolites. While normalizing the data, statistical outliers were scientifically eliminated. Then, a dataset for these non-parametric outliers was created. This dataset was analyzed separately for each variable using Kruskal-Wallis one-way ANOVA test. Prior to Kruskal-Wallis test, for each variable, maximum and minimum IDs from normalized portion of the dataset were incorporated with the outlier dataset for reference. This rescue technique helped in identifying Wilcoxon ranks for the statistical outliers. Furthermore, true outliers in each variable can be identified with proper stringent error checks. True outliers are the potential source of innovation not only in life sciences but also in other branches of science, technology and business.

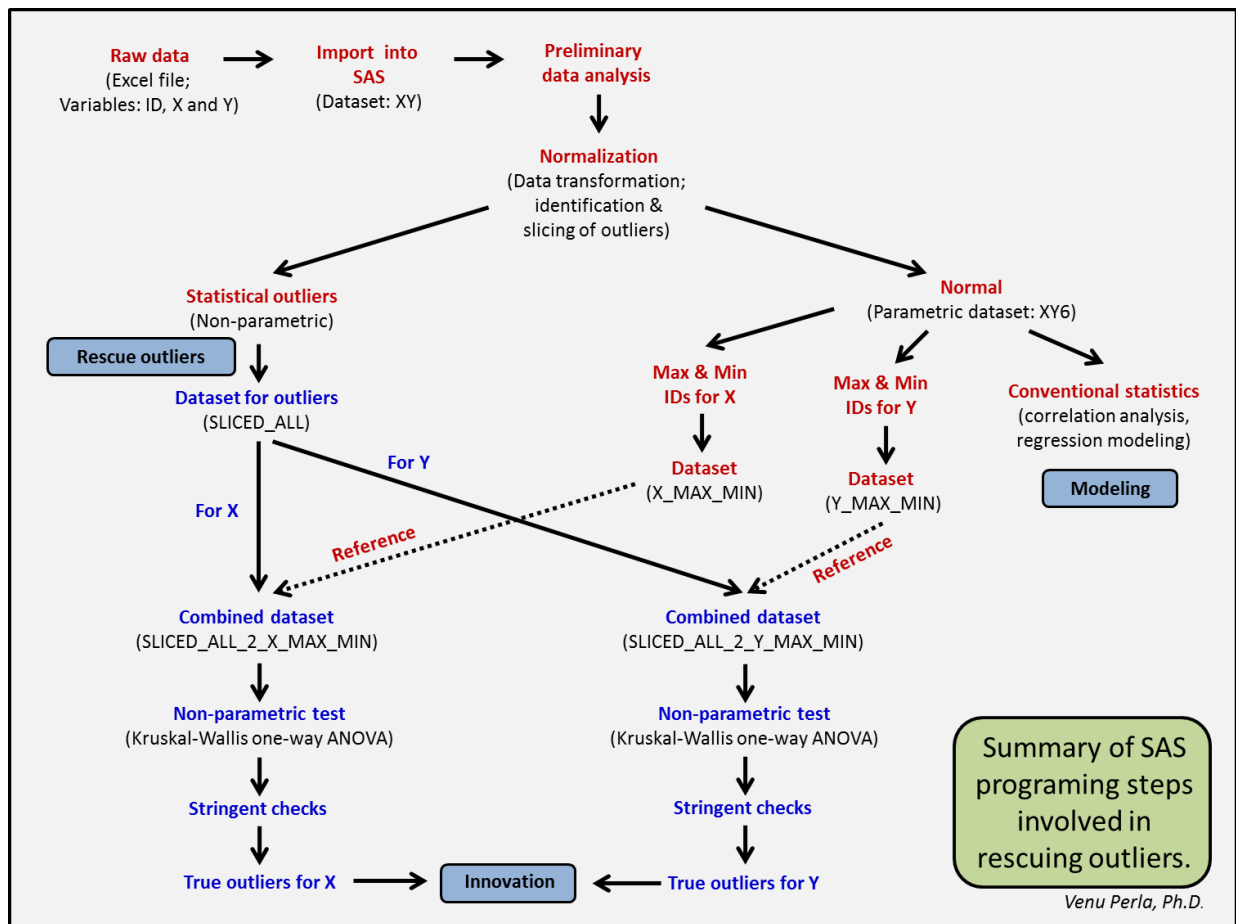


Figure 13

References

Carpenter, Art. 2004. *Carpenter's Complete Guide to the SAS® Macro Language*, Second Edition, SAS® Institute Inc., Cary, NC, USA.

Lafler, Kirk Paul. 2013. *PROC SQL: Beyond the Basics Using SAS®*, Second Edition, SAS® Institute Inc., Cary, NC, USA.

Perla, Venu. 2015. How PROC SQL and SAS® Macro Programming Made My Statistical Analysis Easy? A Case Study on Linear Regression. Ohio SAS® Users Conference held on June 1, 2015 at the Kingsgate Marriott Conference Center at the University of Cincinnati, Cincinnati, Ohio, USA. Available at <http://www.cinsug.org/sites/g/files/g1233521/f/201506/Venu%20Perla%20How%20PROC%20SQL%20and%20SAS%C2%AB%20Macro%20Programming%20Made%20My%20Statistical%20Analysis%20Easy%20A%20Case%20Study%20on%20Linear%20Regression.pdf>

SAS® 9.4 Product Documentation, SAS Institute Inc., Cary, NC, USA. Available at <http://support.sas.com/documentation/94/index.html>

SAS/STAT® 9.3 User's Guide, SAS Institute Inc., Cary, NC, USA. Available at http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#intro_toc.htm

SAS® 9.2 Macro Language: Reference, SAS Institute Inc., Cary, NC, USA. Available at <http://support.sas.com/documentation/cdl/en/mcrolref/61885/HTML/default/viewer.htm#titlepage.htm>

SAS® 9.3 SQL Procedure User's Guide, SAS Institute Inc., Cary, NC, USA. Available at <http://support.sas.com/documentation/cdl/en/sqlproc/63043/HTML/default/viewer.htm#titlepage.htm>

Wikipedia. Outliers. Available at <https://en.wikipedia.org/wiki/Outlier>

Acknowledgments

I would like to thank the organizers for giving me an opportunity to present this paper at Philadelphia SAS® Users Group Fall Meeting on October 29, 2015 at Penn State Great Valley School of Graduate Professional Studies, Malvern, PA, USA. Also, I would like to thank Mr. Surya Perla for proofreading this article.

Trademark Citations

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.

Author Biography



Venu Perla Ph.D. is a biomedical researcher with about 14 years of research and teaching experience in an academic environment. He is currently working in West Virginia. He served the Purdue University, Oregon Health & Science University, Colorado State University, Kerala Agricultural University (India) and Mangalayatan University (India) at different capacities. Dr. Perla has published 13 peer reviewed research papers and 2 book chapters, obtained 1 international patent (on orthopaedic implant device), gave 8 talks and presented 18 posters at national and international scientific conferences in his professional career. Dr. Perla was invited to serve as an editorial board member for several national and international scientific journals. He was trained in clinical trials and clinical data management. He was also trained in advanced SAS® programming and clinical biostatistics at the University of California, San Diego. Currently, he is actively

employing SAS® programming techniques in his research data analysis.

Contact Information

Phone (Cell): (304) 545-5705
 Email: venuperla@yahoo.com
 LinkedIn: <https://www.linkedin.com/pub/venu-perla/2a/700/468>

Appendix

Table 1. SHEET1 of DATAXY.XLS (Microsoft Excel 97-2003 file).

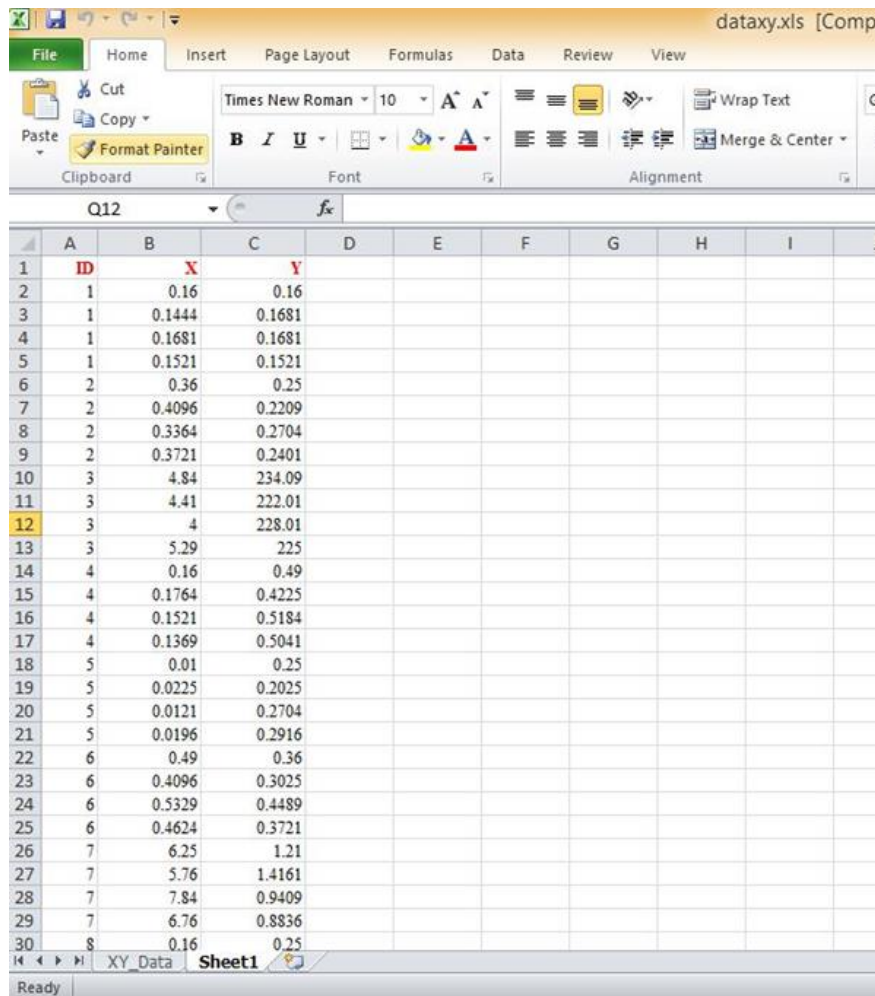
ID	X	Y
1	0.16	0.16
1	0.1444	0.1681
1	0.1681	0.1681
1	0.1521	0.1521
2	0.36	0.25
2	0.4096	0.2209

2	0.3364	0.2704
2	0.3721	0.2401
3	4.84	234.09
3	4.41	222.01
3	4	228.01
3	5.29	225
4	0.16	0.49
4	0.1764	0.4225
4	0.1521	0.5184
4	0.1369	0.5041
5	0.01	0.25
5	0.0225	0.2025
5	0.0121	0.2704
5	0.0196	0.2916
6	0.49	0.36
6	0.4096	0.3025
6	0.5329	0.4489
6	0.4624	0.3721
7	6.25	1.21
7	5.76	1.4161
7	7.84	0.9409
7	6.76	0.8836
8	0.16	0.25
8	0.16	0.2116
8	0.1764	0.2704
8	0.1369	0.2401
9	0.25	0.36
9	0.3025	0.3969
9	0.3844	0.3481
9	0.2304	0.3844
10	1.69	0.81
10	1.44	0.7056
10	1.8496	0.9025
10	1.6641	0.7744
11	0.16	0.16
11	0.1225	0.16
11	0.1936	0.1764
11	0.1764	0.1296
12	3.24	2.56
12	2.89	2.89
12	3.0276	2.6896
12	3.3489	2.4964
13	0.25	3.24
13	0.25	2.9241
13	0.25	3.3489
13	0.1936	3.1329
14	0.25	0.25
14	0.36	0.2209
14	0.2916	0.2916
14	0.2209	0.3136
15	0.49	0.49
15	0.4096	0.4356
15	0.5476	0.5476
15	0.5625	0.4624
16	0.09	0.49
16	0.0576	0.4356
16	0.1156	0.5184
16	0.1024	0.5476
17	1.96	0.81
17	1.8769	0.7396
17	2.0164	0.8649
17	1.9881	0.8836

18	0.64	0.36
18	0.5929	0.49
18	0.6889	0.3136
18	0.6561	0.3025
19	1.69	1
19	1.5129	0.9409
19	1.7424	1.21
19	1.8496	0.9025
20	0.36	0.36
20	0.3249	0.49
20	0.36	0.36
20	0.36	0.3249
21	1.44	1
21	1.44	1
21	1.44	0.8836
21	1	0.9216
22	4	4.41
22	4	5.76
22	4.84	4
22	3.24	3.8025
23	0.49	0.36
23	0.49	0.3136
23	0.5476	0.3969
23	0.4356	0.4225
24	1.69	1.21
24	1.69	1.21
24	1	0.9216
24	1.96	1.44
25	1.21	1
25	1.21	1
25	1	0.9025
25	0.8836	1.44
26	4	1.69
26	5.29	1.44
26	4	1.69
26	3.8416	1.7956
27	0.36	0.49
27	0.36	0.4096
27	0.2916	0.5329
27	0.4096	0.5476
28	4.41	2.89
28	4	2.7225
28	4.84	2.9929
28	5.29	2.9584
29	3.24	1.96
29	3.4596	2.1609
29	3.2761	1.9044
29	3.2041	1.9881
30	1.44	0.64
30	1.5376	0.7056
30	1.44	0.5929
30	1.3689	0.6561
31	1	0.49
31	0.9216	0.5929
31	1.21	0.4624
31	1.2544	0.5329
32	4.41	2.25
32	4.41	2.56
32	4.41	2.56
32	4	2.2201
33	1.96	1
33	1.8225	1.21

33	2.0164	1.44
33	2.0449	0.9025
34	0.49	0.64
34	0.4096	0.7056
34	0.5184	0.64
34	0.4624	0.6084
35	0.25	0.25
35	0.3364	0.3025
35	0.2401	0.2601
35	0.2601	0.2304
36	0.81	0.49
36	0.8836	0.5184
36	0.7744	0.49
36	0.7396	0.4624
37	1.44	0.25
37	1.44	0.3025
37	1.69	0.2401
37	1.21	0.2304
38	1.21	0.49
38	1.21	0.5476
38	1	0.5329
38	1.69	0.4624
39	6.25	4
39	6.0025	4.41
39	6.4516	3.8809
39	6.3001	4
40	1	0.49
40	1	0.49
40	0.9216	0.5476
40	1.21	0.4624
41	0.81	0.64
41	0.8649	0.6724
41	0.8281	0.7056
41	0.7744	0.5776
42	9	7.29
42	9	6.25
42	8.2944	7.8961
42	9.61	7.6176
43	17.64	2.25
43	18.49	2.2801
43	17.0569	2.1609
43	18.8356	2.4025
44	0.81	1
44	0.7569	1
44	0.81	1
44	0.81	1
45	3.61	2.56
45	3.8809	2.89
45	3.5344	2.4649
45	3.6481	2.6244
46	1	0.64
46	1	0.6724
46	1	0.6084
46	0.9216	0.5776
47	1.44	0.49
47	1.44	0.5329
47	1.2321	0.4489
47	1.7161	0.5329
48	0.64	0.49
48	0.64	0.5476
48	0.6724	0.4356
48	0.6084	0.4761

49	1.96	0.64
49	2.0449	0.6561
49	1.8769	0.5776
49	2.0449	0.6889
50	1.96	1.96
50	1.96	1.96
50	1.9881	1.9044
50	1.9044	2.0164
51	1	1
51	1.2321	1.0201
51	0.9409	0.9216
51	1.0201	0.9801



Excel SHEET1 Snapshot.

Macro 'TRANSFORM_LAMBDA_2':

```

/*****
transform_lambda_2: A SAS macro for transforming Y-values using lambda value.
All possible power transformations are performed here using lambda value.
    
```

Author: Venu Perla, Ph.D.
 Date: July, 2015.

```

%let othervars = Comma separated list of variables to be included in
                  trans_dataset. This global variable should be created before
                  calling the macro '%TRANSFORM_LAMBDA_2'.
pre_trans_dataset = Name of the dataset with non-zero and non-negative x- and y-
                    values.
Xvar = Name of the X-variable.
Yvar = Name of the Y-variable.
trans_dataset = Name of the dataset for storing transformed data.
*****
%let othervars=          ; **do not forget comma between the variable names;
%macro transform_lambda_2 (pre_trans_dataset= ,yvar= ,trans_dataset= );
  title "Transformation of &yvar.-variable with convenient lambda";
  proc sql;
    create table &trans_dataset as
    select
      &othervars, &yvar,
      1/(&yvar**2) as neg_2_&yvar,
      1/(&yvar**1) as neg_1_&yvar,
      1/(sqrt(&yvar)) as neg_half_&yvar,
      log(&yvar) as zero_&yvar,
      sqrt(&yvar) as half_&yvar,
      &yvar**1 as one_&yvar,
      &yvar**2 as two_&yvar
    from &pre_trans_dataset;
  quit;

  proc print data=&trans_dataset;
  run;
%mend transform_lambda_2;

*%let othervars=          ;
*%transform_lambda_2 (pre_trans_dataset= ,yvar= ,trans_dataset= );
/*****

```

Macro 'REGRESSION_WOUT_OUTLIERS_2':

```

/*****
regression_wout_outliers_2: A master macro for identification and elimination
outliers, and regression analysis of outlier-free data.

```

Author: Venu Perla, Ph.D.
Date: July, 2015

Note: There are 4 macros in this macro. macro 'SLICE_OBS_2' is specific to this master macro.

```

Dataset = Name of the dataset to be used regression analysis.
Indata = It is same as DATASET.
sliced_data = Name of the dataset for storing outlier observations.
outdata = Name of the dataset for storing data after removing outlier
          observations.
Xvar = Name of the x-variable to be used for analysis.
Yvar = Name of the y-variable to be used for analysis.

%let obset = ; Use this global macro variable to create a set of outlier
              observations before running the macro.
&obset = Contains a set of observation numbers (not IDs) to be deleted from
         INDATA. Observations are separated by comma.
*****
%macro regression_wout_outliers_2 (dataset= , indata= , sliced_data= , outdata= ,xvar=
, yvar=);
  %reg_normality (dataset=&dataset, xvar=&xvar, yvar=&yvar);
  %outlier_obs (indata=&indata, xvar=&xvar, yvar=&yvar);

```

```

    %slice_obs_2 (indata=&indata, sliced_data=&sliced_data);
    %no_outlier_data (indata=&indata, sliced_data=&sliced_data, outdata=&outdata);
%mend regression_wout_outliers_2;
/*****

```

Macro 'SLICE_OBS_2':

```

/*****
slice_obs_2: A SAS Macro for deleting outlier/influencing observations from the
dataset.

```

Author: Venu Perla, Ph.D.
Date: July, 2015.

It utilizes OBSET= , an external global macro variable. One can assign more than one observations for macro variable 'OBSET='.

indata = Name of the dataset to be used for slicing observations.
sliced_data = Name of the output dataset for storing only the outlier observations.

%let obset = Set of outlier observation number(s) separated by comma (observations to be removed from INDATA).

Note: obset= is set to zero initially. When there are no observations to be sliced, then it will not produce any sliced observations. This will not affect further data processing in other macros.

```

*****/

%macro slice_obs_2 (indata= ,sliced_data= );
  title "Dataset for outlier observation(s): &sliced_data";
  data &sliced_data;
    do slice=&obset;
      set &indata point=slice;
      output;
    end;
  stop;
run;

  proc print data=&sliced_data;
  run;
%mend slice_obs_2;
/*****

```