DOCUMENT RESUME

ED 448 189 TM 032 150

AUTHOR Wiggins, Bettie Caroline
TITLE Detecting and Dealing with Outliers in Univariate and Multivariate Contexts.
PUB DATE 2000-11-00
NOTE 33p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (28th, Bowling Green, KY, November 15-17, 2000).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Multivariate Analysis; Social Science Research
IDENTIFIERS *Outliers; Statistical Package for the Social Sciences

ABSTRACT

Because multivariate statistics are increasing in popularity with social science researchers, the challenge of detecting multivariate outliers warrants attention. Outliers are defined as cases which, in regression analyses, generally lie more than three standard deviations from Yhat and therefore distort statistics. There are, however, some outliers that do not distort statistics when they are on the mean of Yhat lines. In univariate analyses, finding outliers can be accomplished using Casewise Diagnostics in the Statistical Package for the Social Sciences (SPSS) version 9.0, which as a three standard deviation default that can be changed easily by the researcher. In bivariate and multivariate analyses, finding outliers more than three standard deviations from Yhat is not as easy. Casewise Diagnostics will detect outliers of "Y"; however, in multivariate analyses, statistics can be distorted by a case lying within the arbitrary three standard deviations because it is said to be exerting so much influence or leverage on the regression line that, in fact, the regression line is distorted. There are two popular ways of detecting this leverage, through distance and influence calculations. The most popular statistic for detecting outliers using distance calculations is Mahalanobis. Several other ways of detecting leverage in multivariate cases are available in SPSS 9.0. Once a researcher has identified a case as being a possible outlier, then the choices are to find out if there has been an error in recording the data, or if the outlier is truly an outlier. it can be argued that there are always going to be outliers in the population as a whole, and this is an argument for keeping the score, because it reflects something natural about the general population. If the researcher decides to drop the case, then the researcher should report it and offer reasons why. (Contains 10 figures, 3 tables, and 20 references.) (Author/SLD)

Running head: OUTLIERS IN UNIVARIATE AND MULTIVARIATE CONTEXTS

Detecting and Dealing with Outliers in Univariate and Multivariate Contexts

Bettie Caroline Wiggins

University of Southern Mississippi

Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY, November 15-17, 2000.

Abstract

Because multivariate analyses are increasing in popularity with social science researchers, the challenge of detecting multivariate outliers warrants investigation. Outliers are defined as cases which, in regression analyses, generally lie over three standard deviations from Yhat and therefore distort statistics. There are, however, some outliers that do not distort statistics when they are on the mean or Yhat lines.

In univariate analyses, finding outliers can be accomplished using Casewise Diagnostics in SPSS 9.0, which has a three standard deviation default. This default can be easily changed by the researcher.

In bivariate and multivariate analyses, finding outliers over three standard deviations from Yhat is not as easy. Casewise Diagnostics will detect outliers of Y; however, in multivariate analyses, an statistics can be distorted by a case lying within the arbitrary three standard deviations because it is said to be exerting so much influence or leverage on the regression line that, in fact, the regression line is distorted. There are two popular ways of detecting this leverage, through distance and influence calculations. The most popular statistic for detecting outliers using distance calculations is

Mahalanobis. Additionally, another statistics using distance calculations for detecting leverage in multivariate cases is Cook's. Typical influence statistics are DfBeta(s), Standardized DfBeta(s), DfFit, Standardized DfFit, and Covariance ratio. All of these can be calculated using SPSS 9.0/Analyze,Regression/Linear/Save/. If SPSS 9.0 detects a multivariate outlier using Mahalanobis distance, SPSS 9.0 will print out a Casewise Diagnostics table labeling the potential outlier case number.

Once a researcher has identified a case as being a possible outlier, then choices are simply to find out if there has been an error in recording the data--which can be fixed--or if the outlier is truly an outlier. If truly an outlier, the researcher must decide whether to keep it or delete it. Arguably there are always going to be outliers in the general population as a whole (or people who score beyond three standard deviations). This is an argument for retaining the score--because it reflects something natural in the general population. If the researcher decides to drop the offending case, then the researcher should report it and offer reasons why.

Detecting and Dealing with Outliers in Univariate and Multivariate Contexts

Multivariate analyses have been increasingly vital to social science research in the past few years (Emmons, Stallings, & Layne, 1990; Grimm & Yarnold, 1995; Henson, 199), mainly because multivariate analyses "more closely reflect the reality that most researchers intend to study" and the "methods [help] avoid the inflation of experimentwise Type I error rates" (Fish, 1998; Henson, 1999, pp. 193-194). Along with the increased usage of multivariate methods has come the challenge associated with multivariate outlier detection.

According to Barnett and Lewis (1978), the detection of univariate outliers should be the first step in the detection of multivariate outliers. Whether doing univariate or multivariate analyses, outliers are those cases (data points) which distort statistics (Tabachnick & Fidell, 1996, p. 65). Tabachnick and Fidell stated that univariate outliers have "an extreme value on one variable" and "multivariate outliers are cases with an unusual combination of scores on two or more variables" (p. 66). A false outlier is a data point, which occurs naturally in the population but is identified by the statistic as an outlier (Jarrell, 1994; Iglewicz & Hoaglin, 1993).

Outliers can be found among dichotomous and continuous variables, among both independent and dependent variables, and in both data and results of analyses (Serdahl, 1996). Serdahl (1996) stated that outliers could fall along the X or Y axis and vary in degree in both directions. If an outlier is in the Y direction and has a large residual, it can "potentially influence the regression parameters (i.e., slope and intercept) by pulling the regression line towards the score's Cartesian coordinate so as to minimize the residual error (e) scores" (Serdahl, 1996, p. 7).

Illustrating Univariate Outliers

In univariate regression analyses, researchers popularly designate outliers as cases lying over three standard deviations from either side of the regression line (Yhat) and having large residuals. However, the selection of three standard deviations as the cut-off point is arbitrary. Final decisions regarding outliers is up to the discretion of the researcher as dictated by the subject matter under investigation.

In Figure 1, points A and D lie over three standard deviations from Yhat and are possible outliers. Points B and C fall within three standard deviations from Yhat (as determined by the standard error of estimate) and are not outliers (at least according to common understanding).

Detection of univariate outliers can be easily accomplished using SPSS. If Casewise Diagnostics is selected in SPSS, it will default to three standard deviations. However, the researcher can change this arbitrary default by going into Analyze/Regression/Linear/Statistics/ and clicking on Casewise Diagnostics and entering in the exact number of standard deviations desired.

INSERT FIGURE 1 ABOUT HERE

SPSS 9.0 Casewise Diagnostics states the results as standardized residuals. A standardized residual is a residual divided by an estimate of its standard error. Standardized residuals which are also known as Pearson residuals, have a mean of 0 and a standard deviation of 1. Those data points beyond, for example, three standard deviations (or outliers above n), will be listed in Casewise Diagnostics with residuals beginning with 3.0 (or outliers above n standard deviations).

There are cases or data points that can be outliers on some statistics and not others. Evans (1999) illustrated this and very clearly demonstrated that an outlier on the mean, for example, may not be an outlier on the correlation coefficient. In Table 1, Jessica is clearly an outlier

which distorts the mean of X and Y; however, Jessica's scores do not distort the correlation coefficient.

Some outliers can distort the correlation coefficient, and according to Hecht (1991), Serdahl (1996), and Evans (1999), a greater distortion generally appears in Y-axis (dependent variable) outliers than X-axis (independent variable) outliers. Table 2 presents fictitious raw data illustrating this concept.

INSERT TABLES 1 - 2 ABOUT HERE

In Analysis 1, SPSS/Analyze/Regression/Linear/ Statistics/Casewise Diagnostics was performed on Cases 1 through 20 using the popular default of three standard deviations, and no outliers were found. Figure 2 is the exact SPSS output for Analysis 1 and shows R^2 as .974 and beta as .987.

In Analysis 2, SPSS/Analyze/Regression/Linear/ Statistics/Casewise Diagnostics was performed with Case Number 21 having X=12.5 and Y=20. Visually, Case Number 21 appears to be an outlier on Y but not on X. Figure 3 is the exact SPSS output for Analysis 2 and shows R^2 as .438 and beta as .662. In Figure 3 under Casewise Diagnostics, Case Number 21 has a standardized residual of 4.187. SPSS listed this case because it lies beyond three (SPSS default)

standard deviations from Yhat. Due to the influence of the Y outlier, R^2 has been reduced by 53.6% and beta by .325. According to Evans (1999), the regression equation has been significantly changed due to the change in beta.

In Analysis 3, SPSS/Analyze/Regression/Linear/ Statistics/ Casewise Diagnostics was performed with Case Number 21 having X=20 and Y=12.5. Visually, Case Number 21 appears to be an outlier on X but not on Y. Figure 4 is the exact SPSS output for Analysis 3 and shows R^2 as .422 and beta as .649. However, in Figure 4 SPSS did not produce a Casewise Diagnostics table, indicating that there were no cases lying beyond three standard deviations from Yhat. According to Hecht (1991), X outliers usually influence score variability more than they influence variable relationships. Evans (1999) suggested that the reason Y is considered an outlier and X is not in Analysis 2 and not vice versa in Analysis 3 is "because X is only considered for its impact on Y" (p. 224).

In Analysis 4, SPSS/Analyze/Regression/Linear/ Statistics/ Casewise Diagnostics was performed with Case Number 12 having X=20 and Y=20. Visually, Case Number 21 appears to be an outlier on both X and Y. Figure 5 is the exact SPSS output for Analysis 4 and shows R^2 as.991 and beta as .996. However, in Figure 5 SPSS did not produce a

Casewise Diagnostics table, indicating that there were no cases lying beyond three standard deviations from Yhat. The R^2 and beta are similar to Analysis 2.

INSERT FIGURES 2 - 5 ABOUT HERE

In 1963, Anscombe and Tukey recommended beginning detection of outliers with a scatterplot. SPSS can produce scatterplots very easily by clicking on Analyze/Graphs/ Scatter/Simple/Define and dragging over the Y-axis and X-axis variables. As shown in Figure 6, this type of scatterplot will simply plot the data points, and outliers will usually be apparent because of their distance from other data points. Importantly, the recent APA Task Force on Statistical Inference report also argued for increased graphical examination of data as well as for data screening procedures (Wilkinson & Task Force on Statistical Inference, 1999).

INSERT FIGURE 6 ABOUT HERE

Anscombe and Tukey (1963) also stated that the most important reason for calculating residuals was to detect outliers. SPSS 9.0 (1999) will also produce other types of scatterplots using residuals by clicking SPSS/Analyze/ Linear/Regression/Plots/ and dragging Standardized

Predicted scores (ZPRED) to the X-axis and Standardized Residuals [(Z Scores)(ZRESID)] to the Y-axis. Figure 7 illustrated this plot. When residuals are standardized, interpretation is made much easier (Hoaglin & Welsch, 1978; Iglewicz & Hoaglin, 1993). This scatterplot will plot the standardized residuals, and by looking at the Y-axis, a researcher can immediately see which standardized residuals are over n standard deviations from the mean of 0.

SPSS 9.0 (1999) will produce other types of residual plots. Besides the SPSS/ZPRED/ZRESID plot, Figure 8 illustrates another popular residual plot used by some researchers, which places Studentized deleted residuals (SDRESID) on the Y-axis and DEPENDENT on the X-axis. This may be one of the few times the DEPENDENT goes on the X-axis. It is done this way because of ease of interpretation. SPSS cannot draw a vertical line for the mean, but it can draw a horizontal mean line. Thus, by placing the DEPENDENT on the X-axis, the horizontal line will be drawn from the Y-axis for SDRESIDS. Studentized deleted residuals are defined by SPSS 9.0 (1999) as "the deleted residual for a case that is divided by its standard error. The difference between a Studentized deleted residual and its associated Studentized residual indicates

how much difference eliminating a case makes on its own prediction" (Help Topics).

INSERT FIGURES 7 - 8 ABOUT HERE

However, according to Serdahl (1996), using Studentized Residuals (SRESID) is a better way of detecting outliers because they are more sensitive to both the Y and to some degree X direction outliers. The SRESID involves the calculation of the residual of the data point in question when its influence has been removed from the data regression equation (SPSS 9.0, 1999, Help Topics).

Illustrating Multivariate Outliers

Additionally, multivariate outliers are sought differently between grouped and ungrouped data among continuous variables. Ungrouped data analyses, such as regression, canonical correlation, factor analysis, and structural equation modeling, are used where univariate and multivariate outliers are sought among all cases at once. Grouped data analyses, such as analysis of covariance (ANCOVA), multivariate analysis of variance (MANOVA) or multivariate analysis of covariance (MANCOVA), profile analysis, discriminant function analysis, or logistic regression, are used where univariate and multivariate outliers are sought separately within each group. When

using SPSS 9.0 Mahalanobis distance on grouped data, the researcher will need to make separate runs for each group. In each of these runs, the researcher will have to set up a dummy dependent variable, such as the case number, to find outliers among the set of independent variables.

There are data points that can influence and distort the regression statistic that lie close in distance to the regression line. In this case, the data point would be an outlier with a small standardized residual (Iglewicz & Hoaglin, 1993). Moreover, the data point would be said to be exerting high leverage on Yhat. A potential leverage point is an outlier in the X direction (Serdahl, 1996). Leverage values fall between 0 and 1 (Belsley & Welsch, 1980). Leverage measurements use the centroid (Figure 9), which is located where the mean of X and Y cross at the regression line.

INSERT FIGURE 9 ABOUT HERE

There are two popular ways of detecting this leverage, through distance and influence calculations. The most popular statistic for detecting outliers in multivariate contexts using distance calculations is Mahalanobis distance. SPSS 9.0 (1999) defines Mahalanobis distance as "a measure of how much a case's values on the independent

variables differ from the average of all cases. A large Mahalanobis distance identifies a case as having extreme values on one or more of the independent variables" (Help Topics). Tabachnick and Fidell (1996) stated that "Mahalanobis distance is the distance of a case from the centroid of the remaining cases where the centroid is the point created by the means of all the variables" (p. 67). Henson (1999) stated that "the mean vector is analogous to the mean in the univariate case and the centroid in the bivariate case" and that "the Mahalanobis distance indicates the geometric distance a given case is from the vector of means" (p. 203). Moreover, Henson (1999) stated that "when variables differ in their standard deviations, they do not contribute equally to the geometric distance from the centroid" (p. 204). Stevens (1996) stated that when there is a positive correlation between two variables, the geometric distance is larger and vice versa for a negative correlation. One reason Mahalanobis distance (D^2) is so popular is because it takes into account positive and negative correlations and standardizes the residuals thus avoiding these challenges (Henson, 1999).

Figure 10 is the exact SPSS 9.0 output using raw data from Analysis 2. Mahalanobis distance was calculated using SPSS 9.0/Analyze/Regression/Linear/Save/Distances/

Mahalanobis. When Mahalanobis is selected on SPSS, the Casewise Diagnostics table will print out IF an outlier is found, and it will give the researcher the case number. For those researchers that desire the actual D^2 for each score, SPSS calculates it as a new variable on the SPSS data input window (Table 3). Cases that have extraordinarily large D^2 values may be multivariate outliers.

INSERT FIGURE 10 and TABLE 3 ABOUT HERE

Influence Statistics

If the researcher desires influence statistics, they can be easily selected and are also calculated for each case and automatically placed in the SPSS data input window. Typical influence statistics are DfBeta(s), Standardized DfBeta(s), DfFit, Standardized DfFit, and Covariance ratio. All of these can be calculated using SPSS 9.0/Analyze/Regression/Linear/Save. Influence statistics are defined by SPSS 9.0 as "the change in the regression coefficients (DfBeta(s)) and predicted values (DfFit) that results from the exclusion of a particular case. Standaradized DfBetas and DfFit values (Table 3) are also available along with the covariance ratio, which is the ratio of the determinant of the covariance matrix with a

particular case excluded to the determinant of the covariance matrix with all cases included " (Help Topics).

SPSS 9.0 also calulates Cook's distance. In interpreting Cook's distance, the researcher should look closely at any number greater than one because this would be considered large and a possible outlier. In Table 3, Case Number 21 has a Cook's distance of .59. However, SPSS 9.0 suggests that this case may be an outlier because it printed the Casewise Diagnostics table when Mahalanobis was selected.

When running multivariate analyses or analyses having more than one independent variable, a "plane or hyperplane is generated as opposed to a line of best fit; however, the basic calculations regarding the residual scores are the same" (Serdahl, 1996). To simplify interpretation, multiple independent variables' residuals can be plotted against each independent variable separately.

Determining What To Do With Outliers

Once an outlier has been identified, the researcher must decide what to do with the outlier. Most outliers are caused by the entering of wrong data by either the respondent or the researcher. However, according to Gaussian distribution, it is normal to have about 5% outliers beyond two standard deviations from the mean in

both directions because this mirrors the population at large. More precisely, because the normal distribution line is asymptotic and does not touch the X-axis, there is the realistic possibility of legitimate extreme values. However, very few data points will normally fall in the tails (Hecht, 1991; Evans, 1999).

Additionally, some outliers are not damaging to the regression equation (Hecht, 1991) as illustrated in Table 1. The researcher can easily determine if an outlier is damaging to the regression equation by simply running the equation with and without the outlier (Hecht, 1991; Hoaglin & Welsch, 1978). Hoaglin & Welsch (1978) recommended the researcher compare any changes in beta weights on the two runs. Evans (1999) suggested that, in addition to comparing beta weights, the researcher might want to investigate if the outlier data points were, in fact, honest answers. Honest answers means that the participant did not mistakenly or intentionally fill in incorrect data. According to Evans (1999), if the answers were dishonest, then they should be dropped. Anscombe (1960) and Iglewicz and Hoaglin (1992) stated that one of the most common sources of outliers was measurement or recording errors, and researchers had three choices as to what do to with an outlier; namely, retain, eliminate, or recode the data

point (Wood, 1983; Chatterjee & Hadi, 1988). The inability of any researcher to determine what caused an outlier creates those hard decisions as to whether to retain, recode or drop the outlying data point. Iglewicz and Hoaglin (1993) stated that if the researcher cannot determine the cause of the outlier, it should be used in the data analysis, "even when they appear to be in error" (p. 8). Care should be taken when deciding to eliminate or recode data values. Sufficient rationale should be present to warrant such changes.

REFERENCES

Anscombe, F.J. (1960). Rejection of outliers. Technometrics, 2, 123-147.

Anscombe, F.J., & Tukey, J.W. (1963). The examination and analysis of residuals. Technometrics, 5, 141-160.

Barnett, V., & Lewis, L. (1978). Outliers in statistical data. Chichester, United Kingdom: John Wiley & Sons.

Belsey, P.A., & Welsch, R.E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York: John Wiley & Sons.

Chatterjee, S., & Hadi, A.S. (1988). Sensitivity analysis in linear rgression. New York: John Wiley & Sons.

Emmons, N.J., Stallings, W.M., & Layne, B.H. (1990, April). Statistical methods used in American Educational Research Journal, Journal of Educational Psychology, and Sociology of Education from 1972 through 1987. Paper presented at the annual meeting of the American Educational Research Association, Boston. (ERIC Document Reproduction Service No. ED 319 797)

Evans, V.P. (1999). Strategies for detecting outliers in regression analysis: An introductory primer. In B. Thompson (Ed.), Advances in social science methodology: (Vol. 5, pp. 213-233). Stamford, CT: JAI Press.

Fish, L. (1988). Why multivariate methods are usually vital. Measurement and Evaluation in Counseling and Development, 21, 130-137.

Grimm, L.G., & Yarnold, P.R. (Eds.). (1995). Reading and understanding multivariate statistics. Washington, DC: American Psychological Association.

Hecht, J.B. (1991, April). Least-squares linear regression and Schrodinger's cat: Perspectives on the analysis of regression residuals. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 333 020)

Henson, R.K. (1999). Multivariate normality, what is it and how is it assessed? In B. Thompson (Ed.), Advances in social science methodology: (Vol. 5, pp. 193-211). Stamford, CT: JAI Press.

Hoaglin, D.C., & Welsch, R.E. (1978). The hat matrix in regression and ANOVA. The American Statistician, 32(1), 17-22.

Iglewicz, B., & Hoaglin, D.C. (1993). How to detect and handle outliers. Milwaukee, WI: ASQC Quality Press.

Jarrell, M.G. (1994). A comparison of two procedures, the Mahalanobis distance and the Andrews-Pregibon

statistic, for identifying multivariate outliers. Research in the Schools, 1(1), 49-58.

Serdahl, E. (1996, January). An introduction to graphical analysis of residual scores and outlier detection in bivariate least squares regression analysis. Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, LA.(ERIC Document Reproduction Service No. ED 395 949)

SPSS Graduate Pack 9.0 for Windows (1999). Statistical packages for the social sciences 9.0 for windows. Chicago, IL: SPSS, Inc.

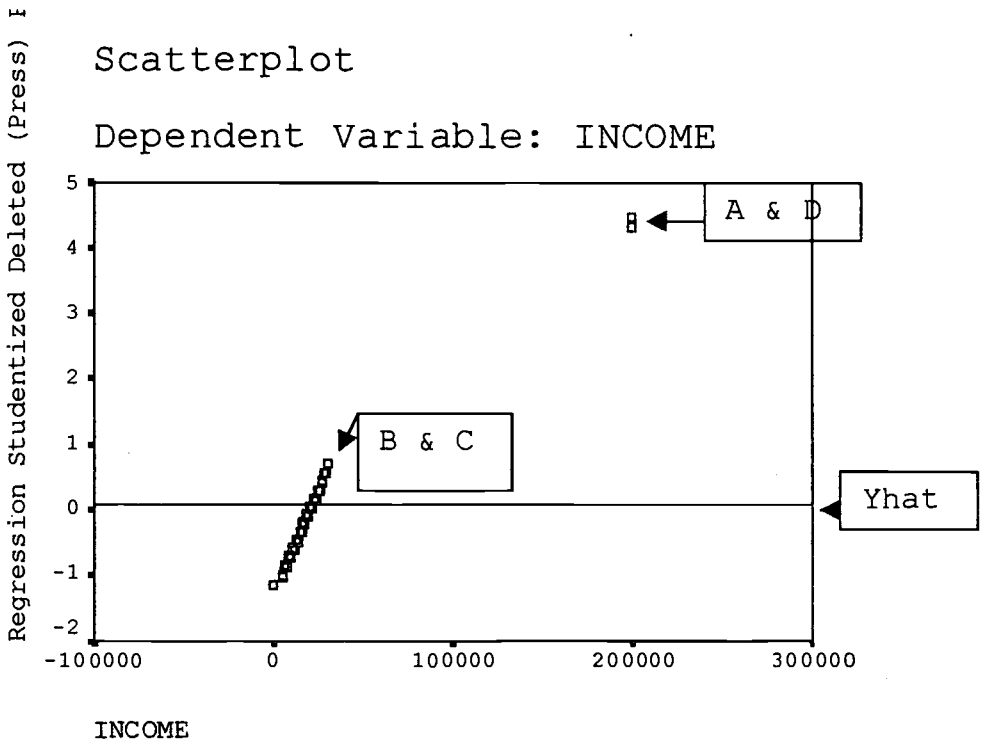
Stevens, J. (1996). Applied multivariate statistics for the social sciences (3rd ed.). Mahwah, NJ: Erlbaum.

Tabachnick, B.G., & Fidell, L.S. (1996). Using multivariate statistics (3rd ed.). New York: HarperCollins College Publishers.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanation. American Psychologist, 54, 594-604.

Wood, F.S. (1983). Measurements of observations far-out in influence and/or factor space. Paper presented at the Econometrics and Statistics Colloquium at the Chicago Graduate School of Business, Chicago, IL.

Figure 1



Casewise Diagnostics

Case Number	Std. Residual	INCOME	Predicted Value	Residual
1	3.008	200000	62833.58153	137166.41847
3	3.096	200000	58820.25767	141179.74233

a Dependent Variable: INCOME

Table 1

	X	Y
Mary	2	2
Robert	4	3
David	6	4
Jessica	90	47

Correlations

		Y	X
Pearson Correlation	Y	1.000	1.000
	X	1.000	1.000
Sig. (1-tailed)	Y	.	.000
	X	.000	.
N	Y	4	4
	X	4	4

Table 2

ID	X	Y
1.000	10.000	10.000
2.000	10.250	10.250
3.000	11.000	11.000
4.000	11.250	11.500
5.000	11.250	11.500
6.000	11.250	11.500
7.000	11.500	11.750
8.000	11.500	11.750
9.000	11.250	11.500
10.000	10.000	10.750
11.000	11.000	11.000
12.000	11.450	12.000
13.000	11.450	11.750
14.000	11.000	11.500
15.000	12.400	12.500
16.000	13.500	13.500
17.000	13.000	13.000
18.000	13.000	13.000
19.000	13.750	14.000
20.000	15.000	15.000

Figure 2. SPSS output for Analysis 1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.987	**.974**	.973	.20466	.974	685.919	1	18	.000

a Predictors: (Constant), X
b Dependent Variable: Y

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.906	.424		2.138	.047
	X	.940	.036	**.987**	26.190	.000

a Dependent Variable: Y

Figure 3. SPSS output for Analysis 2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
Model					R Square Change	F Change	df1	df2	Sig. F Change
1	.662	**.438**	.409	1.64358	.438	14.832	1	19	.001
1	.662	**.438**	.409	1.64358	.438	14.832	1	19	.001

a Predictors: (Constant), X
b Dependent Variable: Y

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
Model		B	Std. Error	Beta		
1	(Constant)	-.637	3.384		-.188	.853
1	(Constant)	-.637	3.384		-.188	.853
	X	1.100	.286	**.662**	3.851	.001
	X	1.100	.286	**.662**	3.851	.001

a Dependent Variable: Y

Casewise Diagnostics

Case Number	Std. Residual	Y	Predicted Value	Residual
21	**4.187**	20.000	13.11792	6.88208
21	**4.187**	20.000	13.11792	6.88208

a Dependent Variable: Y

Figure 4. SPSS output for Analysis 3

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
Model					R Square Change	F Change	df1	df2	Sig. F Change
	.649	**.422**	.391	.95211	.422	13.857	1	19	.001
	.649	**.422**	.391	.95211	.422	13.857	1	19	.001

a Predictors: (Constant), X

b Dependent Variable: Y

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
1	(Constant)	7.609	1.188		6.404	.000
1	(Constant)	7.609	1.188		6.404	.000
	X	.359	.096	**.649**	3.723	.001
	X	.359	.096	**.649**	3.723	.001

a Dependent Variable: Y

Figure 5. SPSS output for Analysis 4

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
Model					R Square Change	F Change	df1	df2	Sig. F Change
1	.996	.991	**.991**	.20297	.991	2199.542	1	19	.000
1	.996	.991	**.991**	.20297	.991	2199.542	1	19	.000

a Predictors: (Constant), X

b Dependent Variable: Y

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
Model		B	Std. Error	Beta		
	(Constant)	.626	.253		2.470	.023
	(Constant)	.626	.253		2.470	.023
	X	.964	.021	**.996**	46.899	.000
	X	.964	.021	**.996**	46.899	.000

a Dependent Variable: Y

Figure 6. Example of Scatterplot with Outlier.

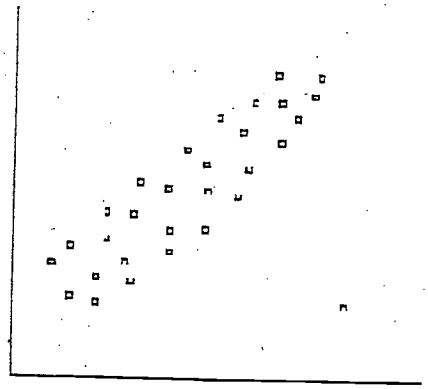
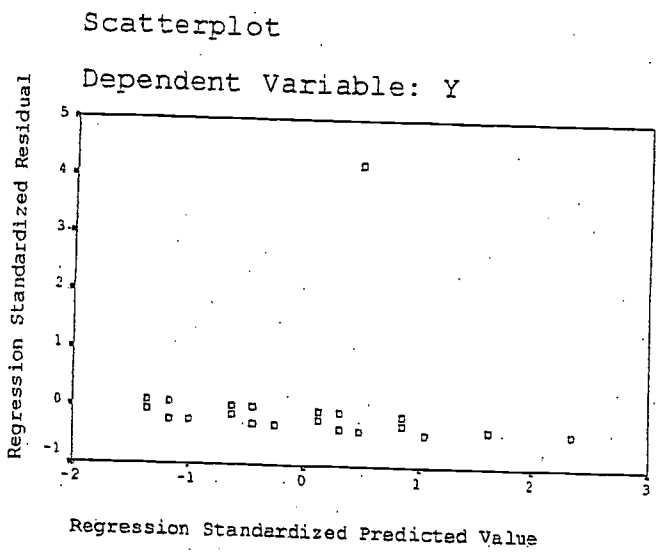


Figure 7. Scatterplot of Dependent Variable Y from SPSS



Note. Regression Standardized Predicted Value = ZPRED; Regression Standardized Residuals = ZRESID. From SPSS 9.0.

Figure 8. Scatterplot using Studentized Deleted Residuals

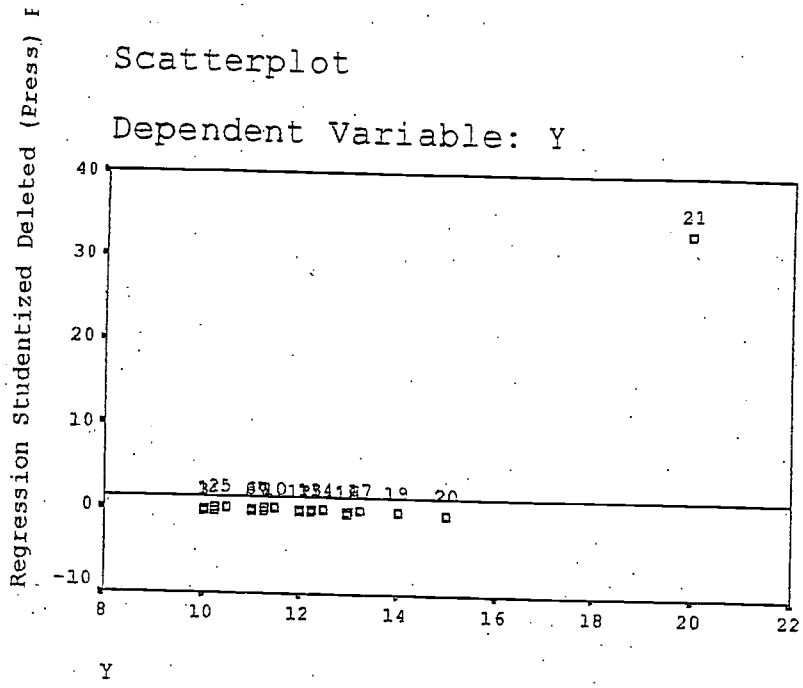


Figure 9. The centroid.

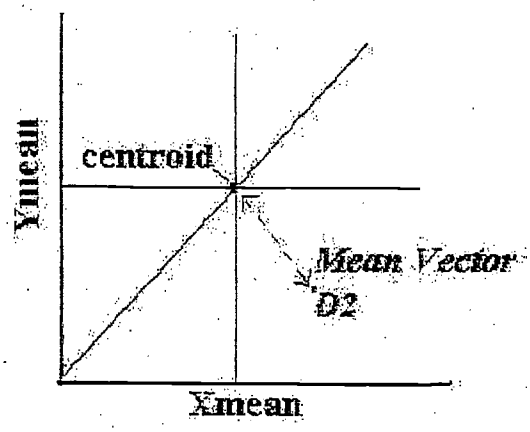


Figure 10. Output for Mahalanobis Distance for Analysis 2

Residuals Statistics

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	10.36688	15.86896	12.32143	1.41540	21
Std. Predicted Value	-1.381	2.506	.000	1.000	21
Standard Error of Predicted Value	.36724	.98850	.48506	.15194	21
Adjusted Predicted Value	10.30297	16.36141	12.35074	1.48684	21
Residual	-.86896	6.88208	-8.45884E-16	1.60197	21
Std. Residual	-.529	4.187	.000	.975	21
Stud. Residual	-.662	4.327	-.008	1.012	21
Deleted Residual	-1.36141	7.34835	-2.93120E-02	1.72930	21
Stud. Deleted Residual	-.652	34.747	1.445	7.633	21
Mahal. Distance	.046	6.282	.952	1.428	21
Cook's Distance	.000	.634	.040	.139	21
Centered Leverage Value	.002	.314	.048	.071	21

a Dependent Variable: Y

Casewise Diagnostics

Case Number	Std. Residual	Y	Predicted Value	Residual
21	**4.187**	20.000	13.11792	6.88208
21	**4.187**	20.000	13.11792	6.88208

a Dependent Variable: Y

Table 3. SPSS Data Input for Analysis 2 Raw Data

ID	D^2	Cook's	Leverage	CovRatio	DEFIT
1.	1.90694	.00485	.09535	1.29212	-.06120
2.	1.40791	.00431	.07040	1.25472	-.05245
3.	.36416	.00305	.01821	1.18187	-.03293
4.	.16736	.00068	.00837	1.17742	-.01438
5.	.16736	.00068	.00837	1.17742	-.01438
6.	.16736	.00068	.00837	1.17742	-.01438
7.	.04611	.00073	.00231	1.16931	-.01406
8.	.04611	.00073	.00231	1.16931	-.01406
9.	.16736	.00068	.00837	1.17742	-.01438
10.	1.90694	.00529	.09535	1.29140	.06391
11.	.36416	.00305	.01821	1.18187	-.03293
12.	.06431	.00001	.00322	1.17380	.00201
13.	.06431	.00047	.00322	1.17170	-.01138
14.	.36416	.00001	.01821	1.19266	.00230
15.	.23521	.00320	.01176	1.17191	-.03206
16.	1.79613	.01764	.08981	1.26178	-.11445
17.	.90529	.00933	.04526	1.20484	-.06841
18.	.90529	.00933	.04526	1.20484	-.06841
19.	2.35488	.01070	.11774	1.31982	-.09776
20.	6.28199	.12409	.31410	1.66608	-.49245
21.	**.31667**	**.63419**	**.01583**	**.00026**	**.46627**

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM032150

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Detecting and Dealing with Outliers in Univariate and Multivariate Contexts

Author(s): Bettie Caroline Wiggins

Corporate Source: University of Southern Mississippi

Publication Date: November 16, 2000

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ↑ ☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ↑ ☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ↑ ☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Bettie C. Wiggins, M.D.	Printed Name/Position/Title: Bettie C. Wiggins, Asst. Prof.
Organization/Address: USM Box 5027 Hattiesburg, MS 39406	Telephone: 601-266-4580 FAX:
	E-Mail Address: bettie.barrett@usm.edu Date: 11/16/00

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com