A Comparison of Three Data Analysis Approaches When Outliers are Suspected

Stephen F. Olejnik
University of Florida

Obtaining a descriptive summary (i.e., group means, variances, minimum and maximum values, etc.) of a data set before statistical tests of hypotheses are conducted is standard operating procedure for most data analysts. The purpose of this preliminary analysis is to identify possible errors in the data set and determine whether basic assumptions of hypothesis tests have been met. Data points are often found which are clearly in error, e.g., a score on a test may exceed the total number of items on the test. These errors are relatively common and are generally the result of scoring, coding or keypunching errors. If they are identified, they can be corrected or deleted if the original data source is no longer available. There are many other situations, however, when one or two data points don't seem to "fit" in with the remaining data set, but no apparent error can be found. These questionable data points are often called outliers.

The data analyst has a choice of several procedures to follow in dealing with outliers. One option is to accept the questionable data points and test the hypotheses of interest using the traditional least squares analysis strategies. While this approach is acceptable, data outliers can have a serious effect on hypothesis tests and parameter estimates. A single outlier can influence results which indicate a treatment main effect or a significant interaction, when the null hypotheses are true. An outlier can also have the opposite effect on an hypothesis test. That is, an outlier can mask a significant main effect or an interaction.

A second option might be to delete the questionable data points and proceed to test the hypotheses of interest. Support for the deletion of data points might be provided through a number of tests for outliers (Barnett and Lewis, 1978). Data analysts, however, are reluctant to delete data unless they are confident that an error had been made and that the analysis without deletion would be meaningless. However, there is no consensus on which statistical test for outliers is most appropriate (Fisher, 1981).

A third alternative to the problem of questionable data points attempts to reduce or eliminate the influence outliers have on parameter estimates and hypothesis tests. Least square estimators, such as a sample mean, are affected greatly by extreme values. That is, in a sample distribution the addition or deletion of an outlier would change a parameter estimate considerably more than the addition or deletion of any other point near the center of the distribution. To reduce or eliminate this disproportional influence on a parameter estimate or hypothesis test, data transformation or differentially weighting can be used. This approach is relatively new and additional research with it is needed. The purpose of the present paper is to consider these three data analysis options in analyzing the results of an experiment believed to contain one or two erroneous data points and to compare the conclusions from each procedure.

The Experiment

The data for this study were obtained from an experiment investigating the relationship between test anxiety, a test taking strategy, and performance on a multiple choice test. (Schmitt & Crocker, 1981) In that study, undergraduate and graduate students enrolled in an introductory measurement class were randomly assigned to one of two groups. Students in group one were given a 15 item multiple choice test on concepts taught during the first several weeks of the course. Students in the second group were asked to write their own answers to the same 15 items before looking at and choosing one of the answer foils provided. The dependent variable for the study was the number of correctly answered multiple choice items. All of the students had completed the Mandler-Sarason Test Anxiety Scale earlier in the quarter. Complete data sets from 73 students were available for analysis.

Least Squares Analysis of Covariance

The data were first examined descriptively by determining the means, standard deviations, minimum and maximum values on both the test anxiety scale and the 15 item multiple choice test. The results of this analysis are reported in Table 1. The results reported in Table 1 do not indicate the presence of any erroneous data points. The range of scores

Table 1

Means, standard deviations minimum and maximum values by group for the test anxiety scale and the 15 item multiple choice test.

		Mandler-Sarason Test Anxiety Scale				15 Item Multiple Choice Test			
	N	Mean	Standard Deviation	Minimum	Maximum	Mean	Standard Deviation	Minimum	Maximum
Group 1	34	15.23	6.71	3	27	10.05	2.08	5	14
Group 2	39	17.79	6.71	4	31	10.25	2.78	3	15

within each group is similar and they lie within the scoring range of each instrument. Furthermore, the means and standard deviations for each measure were similar for the two groups.

The hypothesis of primary interest was that students who were asked to write their own answers to the 15 item test before choosing an answer from those provided would score significantly higher than students who just chose an answer from the options provided. To test this hypothesis the data were analyzed using the least squares one-way analysis of covariance with test anxiety as the covariate. This analysis was computed using the GLM subprogram of the Statistical Analysis System (SAS) computing package. Before testing the hypothesis of interest, the assumption of no covariate by group interaction (homogeneity of regression slopes) was tested. The computed F statistic was 2.29 ($p<.135$) under the hypothesis of no interaction. The group effect was then tested. The computed F was .21 ($p<.646$) under the hypothesis of no difference between groups. Finally, the relationship between anxiety and the dependent variable was tested. The computed F was 9.57($p<.003$) under the hypothesis of no relationship between the two measures. The conclusions drawn from these analyses were that the relationship between test anxiety and the performance on the multiple choice test was not dependent on the treatment received and that the test taking strategy of first answering the question before referring to the options provided was not effective in improving performance on the multiple choice test.

The residuals from the full regression model which included the interaction term were examined as a further check for outlying data points. Draper and Smith (1966) and Kleinbaum and Kupper (1978) both suggest that residuals lying three or more standard deviations away from their mean might be considered as outliers. Both references, however, caution against the deletion of outliers without further investigation. Column 2 of Table 2 presents the residuals for the least squares regression equation.

The standard error of estimate for the model was 2.23. Only one data point exceeded 3 standard errors and that point corresponded to Obs 73. A second data point corresponding to Obs 72 exceeded the mean of the residuals by a factor of 2.9. Both of these points were therefore considered as prime candidates for deletion as outliers. Before they were deleted and the data reanalyzed, however, both studentized residuals and Cook's (1977) distance index were computed and examined. These statistics were obtained using the *9R* subprogram of the BMDP computing package.

Least Squares ANCOVA with Data Deleted

Rather than examining the raw residuals, it is frequently recommended that studentized residuals for the results of the experiment are reported in column 3 of Table 2. Lund (1975) has developed a series of tables containing critical values that can be used with studentized residuals to test the hypothesis that the data set contains no data outliers. In this data set the largest studentized residual was -3.24, associated with the Obs 73. The critical value for the studentized residual based on a sample of 70 individuals at $\alpha = .10$ is 3.11 and 3.29 at $\alpha = .05$. Thus, the hypothesis that the data set does not contain a data outlier would be rejected at the .10 level of significance but not at the .05 level. In this case, it might be argued that a Type II error is more serious than a Type I error and thus the increased alpha level is justified. That is, a data analyst would not want to conclude that there are no outliers when in fact there is an outlier. As pointed out earlier, an undetected outlier can seriously affect both hypothesis tests and parameter estimation.

In addition to the studentized residual, Huber (1975) and Davies and Hutton (1975) have suggested that the variances of the residuals provide useful information on erroneous data points. Cook (1977) developed a procedure which combines the studentized residual with the variance of the residual to provide a single index of the influence each data point has on the estimation of parameters. Cook's distance index is estimated for each observation as follows:

$$D_i = \frac{t_i^2}{p} \frac{V(Y_i)}{V(R_i)}$$

Where t_i is the studentized residual for the ith observation;
is the number of parameters estimated;
$V(Y_i)$ is the variance of the predicted observation;
and $V(R_i)$ is the variance of the residual. (Cook, 1977, p. 16)

Table 2

Raw residuals, studentized residuals, Cook's distance indices and Huber's robust weights associated with each data observation in the experiment.

Obs	Residual	Studentized Residual	Cook's Distance	Robust Weights C=2.0	Robust Weights C=2.5
1	4.6736	2.14	.05	.8282	.9330
2	2.4229	1.10	.01	1.0	1.0
3	.7802	.35	.00	1.0	1.0
4	.9544	.46	.01	1.0	1.0
5	-.4331	-.20	.00	1.0	1.0
6	3.3708	1.61	.09	1.0	1.0
7	1.9589	.89	.01	1.0	1.0
8	.1376	-.06	.00	1.0	1.0
9	-.4554	-.21	.00	1.0	1.0
10	.6036	.27	.00	1.0	1.0
11	2.0724	.98	.03	1.0	1.0
12	4.2442	1.93	.03	.9513	1.0
13	-1.2784	-.58	.00	1.0	1.0
14	-.7472	-.35	.00	1.0	1.0
15	1.3085	.60	.01	1.0	1.0
16	.7082	.33	.00	1.0	1.0
17	1.0167	.47	.00	1.0	1.0
18	1.5669	.74	.01	1.0	1.0
19	-3.1477	-1.45	.03	1.0	1.0
20	-1.5145	-.69	.00	1.0	1.0
21	2.7802	1.26	.01	1.0	1.0
22	-.0065	-.00	.00	1.0	1.0
23	-.1851	-.09	.00	1.0	1.0
24	-4.2194	-1.92	.03	.9116	1.0
25	.1347	.06	.00	1.0	1.0
26	-2.8624	-1.30	.01	1.0	1.0
27	-1.5051	-.69	.00	1.0	1.0
28	-2.3964	-1.09	.01	1.0	1.0
29	.1722	.08	.00	1.0	1.0
30	1.7216	.78	.00	1.0	1.0
31	-2.3374	-1.06	.01	1.0	1.0
32	1.4229	.65	.00	1.0	1.0
33	-1.3964	-.64	.00	1.0	1.0
34	-1.5145	-.69	.00	1.0	1.0
35	.7802	.35	.00	1.0	1.0
36	-1.0423	-.48	.00	1.0	1.0
37	-1.8686	-.88	.02	1.0	1.0
38	-1.0065	-.47	.00	1.0	1.0
39	-1.8653	-.87	.01	1.0	1.0
40	1.2495	.58	.01	1.0	1.0
41	-2.0411	-.93	.01	1.0	1.0
42	.4855	.22	.00	1.0	1.0
43	1.8869	.86	.01	1.0	1.0
44	.1030	.05	.00	1.0	1.0
45	1.3118	.62	.01	1.0	1.0
46	-1.0411	-.47	.00	1.0	1.0
47	1.2442	.57	.00	1.0	1.0
48	.9544	.46	.01	1.0	1.0
49	.6036	.27	.00	1.0	1.0
50	-2.4331	-1.14	.03	1.0	1.0
51	3.8981	1.78	.03	.9920	1.0
52	-.6915	-.32	.00	1.0	1.0
53	-.6292	-.30	.00	1.0	1.0
54	-.3638	-.17	.00	1.0	1.0
55	-2.7904	-1.29	.03	1.0	1.0
56	2.6016	1.18	.01	1.0	1.0
57	1.7802	.81	.00	1.0	1.0
58	2.1376	.97	.01	1.0	1.0
59	-1.0065	-.47	.00	1.0	1.0
60	-.9833	-.45	.00	1.0	1.0
61	1.7802	.81	.00	1.0	1.0
62	2.6016	1.18	.01	1.0	1.0
63	.8987	.41	.00	1.0	1.0
64	2.6036	1.19	.01	1.0	1.0
65	-1.5735	-.72	.01	1.0	1.0
66	1.3085	.60	.01	1.0	1.0
67	1.2096	.56	.01	1.0	1.0
68	.1376	.06	.00	1.0	1.0
69	-1.1131	-.51	.00	1.0	1.0
70	.0167	.01	.00	1.0	1.0
71	-3.9691	-1.83	.05	.9780	1.0
72	-6.5051	-2.97	.08	.5893	.6617
73	-6.7212	-3.24	.42	.5393	.6119

Data observations with large D_i values are identified as outliers. The effect of the removal of the unusual data point is judged in terms of the confidence region for the population parameters. Cook provided an example in which the results of an examination of studentized residuals and his distance statistic provided different conclusions. He concluded that the data point identified as an outlier using studentized residuals did not have a large impact in the estimation of the regression parameter and thus should not be considered as an outlier.

In the experiment being considered here, Cook's distance statistic was calculated for each observation and the results are reported in column 4 of Table 2. The largest distance statistic was obtained for Obs 73. This was the same data point identified as a significant outlier ($p = .10$) using the studentized residuals. Since both procedures identified the same data point as being an outlier, it was deleted and the remaining data set was reanalyzed. Table 3 presents the results of this analysis in comparison with the analysis of covariance without data deletion. The results of the reanalysis produced a computer F ratio of 5.84

Table 3

Summary results of five ANCOVA analyses testing the hypothesis of a group by anxiety interaction.

Procedure	F	p<
ANCOVA (without deletion)	2.29	.135
ANCOVA (with data deleted)	5.84	.018
Robust ANCOVA (c=2.5)	3.61	.062
Robust ANCOVA (c=2.0)	4.10	.047
Nonparametric ANCOVA	$\chi^2 = 2.346$	.126

($p < .018$) for the interaction of test anxiety and group given the hypothesis of no interaction. Hence, it was concluded that the effectiveness of the treatment was dependent on the level of student test anxiety at the beginning of the quarter.

Further analysis using the Johnson-Neyman technique (Kerlinger and Pedhazur, 1973) indicated that writing the answers to the multiple choice test was effective for low anxious individuals but was detrimental for high anxious individuals. With the data point deleted, the new regression coefficients moved to the edge of a 25.86 percent confidence ellipsoid for the regression parameters. These results demonstrate that when data sets contain questionable data points, choosing between data analysis option 1 (non-deletion) and option 2 (deleting data) can provide results and conclusions that are contradictory.

The third alternative being considered here is to transform or weight the data points such that the questionable data do not have a disproportional influence on the hypothesis test or the estimation of the regression parameters. A procedure for differentially weighting observations has been suggested by Huber (1964, 1972) as one approach to a broader group of strategies referred to as robust analysis techniques (Box 1953; Tukey, 1962). A procedure for transforming the data using ranks in a nonparametric analysis of covariance procedure has been suggested by Shirley (1981). These two approaches are considered in the next section as a third analysis strategy when data outliers are suspected.

ANCOVA with Transformed and Differentially Weighted Data

A major problem with methods that identify outliers based on the examination of residuals is that the unusual data points pull the estimated parameters toward them and, as a result, the outliers may not appear different and can easily be overlooked. Statisticians concerned with this problem have developed a new methodology which consists of several procedures which are referred to as ROBUST estimation techniques. They are robust in the sense that the normality assumption can be violated and the procedures still provide valid estimates of the parameters of interest. If the distribution is normal, then the estimates are very similar to the least squares estimates, but when the underlying distribution is nonnormal (i.e., long tail distributions) the estimators can be considerably different. The parameters estimated in this study were maximum likelihood estimates; the procedure used here is sometimes referred to as an M-estimation technique.

The essential feature of the robust procedures is that each observation is given a weight such that extreme values (outliers) are given less value when the parameter is estimated. The point at which observations are considered to be extreme is referred to as the tuning constant and corresponds roughly to the number of standard deviations beyond which the data analyst desires to begin reducing the weight of the observation.

A full discussion of the technical aspects of robust estimation techniques can be found in several books and numerous articles: Launer and Wilkinson, 1979; Hogg, 1979 and Huber, 1972. This study used a weighting function suggested by Huber (1964) as a possible solution to the problem of data outliers in the experiment under consideration. The determination of the actual data weights requires an iterative computer routine. A program to compute robust estimates using the MATRIX subprogram of SAS, written by Pendergast[1], was used in the present paper.

Two tuning constants were considered in analyzing the experimental data set. The first constant was set equal to 2.0 and the second constant was set equal to 2.5. The weights for each data point under the two tuning constants are reported in columns 5 and 6 of Table 2. With the larger tuning constant (C=2.5), only three data points were given weights less than one. These points corresponded to the observations having the largest residuals. The observation having the greatest influence on the parameters as identified by Cook's distance statistic was given the least weight in the robust estimation. The two other observations which were given weights less than 1 were not identified by Cook's procedures as having a large influence in the estimation of the parameter. The test for the interaction of anxiety with the treatment levels using the weighted observations resulted in a computed F statistic equal to 3.611 ($p < .062$). Table 3 presents these results in comparison with the two previous ANCOVA analyses deleting and not deleting Obs 73.

When the smaller tuning constant (C=2.0) was used, seven data points were given weights less than 1. Again, the reduced weights were assigned to the data points having the largest residuals. The smallest weight was assigned to observation 73, the same point identified by Cook's distance as the most influential observation in the estimation of the regression parameters. Using the weighted observations with the 2.0 tuning constant, the test for the interaction resulted in a computed F of 4.10 ($p < .047$). Table 3 reports the comparison of these results with the previous analyses.

A second procedure that might be suggested to modify a data set such that outlying data points do not have a disproportional influence in hypothesis tests or parameter estimation is to transform the data set. One simple transformation which might be suggested is a rank transformation, i.e., observations are rank ordered from the smallest value to the largest value. With this procedure, data are not deleted and all observations in the data set have an equal influence on the hypothesis test and parameter estimate. Such an analysis was suggested by Quade (1967). The rank analysis of covariance procedure suggested by Quade, however, assumed no covariate by treatment interaction and does not provide a procedure for testing this assumption. In the data set being considered herein, some evidence suggests the presence of a test anxiety by group interaction, thus making Quade's procedure inappropriate for the present problem.

A nonparametric approach to analysis of covariance was suggested by Shirley (1981) based on the work by Bennett (1968). Both the covariate and dependent variables are rank ordered independently and the resulting transformed data set is then analyzed using a standard analysis of covariance program. The total mean square for the unadjusted analysis of variance on the ranks is substituted for the usual adjusted mean square error. The resulting statistic has a chi-square distribution with degrees of freedom equal to the number of treatment groups minus one.

Following the procedure suggested by Shirley (1981) the scores on the test anxiety scale and the 15 item multiple choice test scores were rank ordered using the RANK subprogram of the SAS computing package. The ranked data were then analyzed using the GLM subprogram for analysis of covariance in the SAS package. Finally, the computed test statistic was calculated for the interaction of test anxiety with group using the total mean square as the error term. The computed ratio was 2.35 ($p < .126$). These results, compared to the previous NACOVA results are presented in Table 3. Thus, using a rank transformation to modify the data set in order to control for outlying data points resulted in the same conclusion as option 1 (non-deletion).

Discussion

The present paper has considered three options for data analysis options when outliers are suspected. First, the possibility of an outlier can be ignored and traditional least squares procedures can be used. Second, the suspected outlier(s) can be deleted and the remaining data set analyzed using least squares procedures. Third, a procedure which might be considered as a compromise between options one and two, the data set may be transformed or weighted differentially but data are not deleted. When these three options were implemented using a data set obtained from an experimental study, the results and conclusions were contradictory. Data analysis option 1 indicated no significant treatment effect. Option 2 indicated a significant interaction effect with the treatment effective for some but not all subjects. Option 3 provided even more confusing results in that transforming the data using ranks provided conclusions similar to those from option 1; however, by differentially weighting the observations the results were similar to those provided by option 2.

These results clearly demonstrate the difficulty one can encounter when interpreting the results of a study containing possible data outliers. It also demonstrates a weakness of a research study that relies on only one analysis strategy in testing hypotheses of interest without close examination of the distributional properties of the data set being analyzed. Depending on which procedure is chosen, the conclusions from the study could differ. These results can be generalized to the analysis of evaluation data. It is important, therefore, to carefully consider the assumptions implied by the selected analysis strategy and to determine whether the assumptions have been met. Furthermore, a routine descriptive analysis of a given data set should be made in order to identify potentially erroneous data points.

Whenever some doubt is raised regarding the distributional properties of the data set or when analytic assumptions may have been violated, one should consider a multiple analysis approach. If the analyses are in agreement, then the researcher can place greater confidence in the conclusion. However, if the conclusions are contradictory, then the researcher needs to examine the analysis strategies more closely and attempt to explain the contradiction. It might, for example, be a simple question of power, with one strategy being more sensitive to treatment effects than the others. If no explanation can be found, then the researcher should view the results as tentative and a replication of the study should be conducted.

In the case of the experiment that was considered throughout the paper, a replication was conducted. The results of that investigation confirmed a group of test anxiety interaction with the test taking strategy being studied having a positive effect with low anxious individuals. Although a multiple analysis approach might be costly, ignoring this strategy could have detrimental effects for the program being evaluated. A thorough analysis of any data set is critical to the appropriateness of all final conclusions and recommendations.

The recommendation of multiple analysis strategies raises the question of which set of procedures the data analysis should select. Certainly, the researcher should choose procedures which make different assumptions concerning the data set being analyzed. If the strategies provide different conclusions, the differences might be explained by violations in assumptions. Additional research is needed, however, to identify strengths and weaknesses of a variety of analysis procedures under differing conditions. Some work in this areas has been done in the comparisons of parametric with nonparametric procedures (Blair & Higgins, 1980) and in comparisons of parametric with robust analysis procedures (Huber, 1973). Further study is needed and a compilation of the results is recommended.

Further work is also needed in studying the small sample properties of the robust estimation techniques and the nonparametric analysis of covariance procedure considered in this paper. A comparison of Type I error rates and power should be made when treatment groups consist of a small number of subjects and the underlying distribution of the data is non-normal.

Finally, the present paper considered three data analysis options when a single data outlier was suspected. Additional investigations are needed to determine the appropriateness of these options when data sets contain multiple outliers.

References

Barnett, V. and Lewis, T. Outliers in Statistical Data. New York: John Wiley & Sons, 1978.

Behnken, D. W. and Draper, N. R. Residuals and Their Variance Patterns, Technometrics, 1972, 14, 101-111.

Bennett, B. M. Rank-order. Test of Linear Hypothesis, J. Statist. Soc. B, 1968, 30, 483-489.

Blair, R. C. and Higgins, J. J. A Comparison of the Power of Wilcoxon's Rank-Sum Statistic to That of Student's *t* Statistic Under Various Nonnormal Distributions. Journal of Educational Statistics, 1980, 5, 309-336.

Box, G. E. P. Non-Normality and Tests on Variances, Biometrika, 1953, 40, 318-335.

Cook, R. D. Detection of Influential Observation in Linear Regression, Technometrics, 1977, 19, 15-18.

Davies, R. B. and Hutton, B. The Effects of Errors in the Independent Variables in Linear Regression, Biometrika, 1975, 62, 383-391.

Draper, N. and Smith, H. Applied Regression Analysis. New York: John Wiley & Sons, 1966.

Fisher, P. A Comment on the Subjective Decisions Required of the Researchers in the Selection of a Statistical Outlier Test, Florida Journal of Educational Research, 1980, 22, 27-41.

Hogg, R. V. Statistical Robustness: One View of its use in Applications Today, The American Statistician, 1979, 33, 108-115.

Huber, P. J. Robust Estimation of a Location Parameter, Annals of Mathematical Statistics, 1964, 35, 73-101.

Huber, P. J. Robust Statistics: A Review, Annals of Mathematical Statistics, 1972, 43, 1041-1067.

Huber, P. J. Robust Regression: Asymptotic, Conjectures and Monte Carlo. Annals of Statistics, 1973, 1, 799-821.

Huber, P. J. Robustness and Designs. A Survey of Statistical Design and Linear Models. North Holland, Amsterdam, 1975.

Kerlinger, F. N. and Pedhazur, E. J. Multiple Regression in Behavioral Research. New York: Holt Rinehart and Winston, Inc. 1973.

Kleinbaum, D. G. and Kupper, L. L. Applied Regression Analysis and Other Multivariable Methods. North Scituate MA: Duxbury Press, 1978.

Launer, R. L. and Wilkinson, G. M. Robustness in Statistics. New York: Academic Press, 1979.

Lund, R. E. Tables for an Approximate Test for Outliers in Linear Models, Technometrics, 1975, 17, 473-476.

Quade, D. Rank Analysis of Covariance, Journal of American Statistical Association, 1967, 62, 1187-1200.

Schmitt, A. P. and Crocker, L. Improving Examinee Performance on Multiple Choice Tests, Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA, April, 1981.

Shirley, E. A. C. A Distribution-free Method for Analysis of Covariance Based on Ranked Data, Applied Statistics, 1981, 30, 158-162.

Tukey, J. W. The Future of Data Analysis, Annals of Mathematical Statistics, 1962, 33, 1-67.

NOTE

1. The author wishes to thank Dr. Jane Pendergast for her assistance and the use of her program in the determination of the robust estimates and hypotheses tests with those estimates.