Outlier Lies: An Illustrative Example of Identifying Outliers and Applying Robust Models

Karl Ho, University of North Texas
Jimmie R. Naugher, University of North Texas

The presence of outliers can contribute to serious deviance in findings of statistical models. In this study, we illustrate how a minor, typographical error in the data could make a standard OLS model "lie" in the estimates and model fit. We propose robust techniques that are insensitive to extreme, outlying cases and provide better predictions. With implementation examples, we demonstrate how robust technique improves estimations over conventional models based on normality and outlier-free assumptions.

The possibility of outliers is an important consideration when applying regression statistics such as R^2 and the Pearson product moment correlation coefficient (Huber 1981, Hempel *et al* 1986). We provide an example in this article that illustrates how dramatic the influence of only a tiny portion of the data can have on the model estimate and goodness of fit statistics. In the following analysis, we demonstrate that with two outliers included in a data set of 48 observations, only 15% of the variation in the dependent variable is accounted for by the differences on the independent variable ($r = .39$ and $r^2 = .15$, N=48). However, when the two outliers are removed, 48% of the variation is accounted for ($r = .69$ and $r^2 = .48$, N=46).

The data are from a survey of metropolitan colleges and universities conducted by the Office of University Planning at the University of North Texas. The institutions ranged from some with essentially open admissions to those with selective admissions criteria. The independent variable is the *institution's average SAT score for new freshmen* and the dependent variable is the *institution's six-year graduation rate.* As expected, there was a strong linear relationship between the average SAT score for new freshmen and the graduation rates. However, only two outliers can hide this fact in terms of r and r^2 analysis. There are three purposes to this article:

- To illustrate how only two outliers can have a dramatic influence on r and r^2 values.
- To demonstrate that outliers can be identified by visual inspection of the scattergram, provided the difference is extreme enough.
- To point to statistical tools that provide more reliable statistical means to identify outliers than visual inspection alone.

The reported SAT averages ranged from 464 to 1152. The reported graduation rates ranged from 12.0% to 74.4%. The outliers reported the two lowest average SAT scores with relatively high graduation rates, *i.e.*, an SAT of 464 with a graduation rate of 44.1% (near the middle) and an SAT of 598 with a graduation rate of 72.0% (near the top). Institutions were requested to use the total SAT for averages, for which 400 is the lowest possible value. An average SAT of 464 or 598 is not believable. (Probably a clerk recorded either the math SAT or verbal SAT instead of the total SAT. Doubling the two reported SAT values of 464 and 598 yields values that fit well with the graduation rates.)

Figure 1 is based on the 48 cases that include the two outliers. The SAT values and graduation rates are plotted as a graph and the resulting regression line is plotted. Note how the paired values of SAT=464 and graduation rate=44.1 and SAT=598 and graduation rate=72.0 are isolated in the top left corner of the graph. The two points "lie outside" the general pattern formed by the other cases. The R^2 is 0.1523.

Figure 2 is based on 46 cases, with the two outliers excluded. The SAT values and graduation rates as shown in Table 1 are plotted as a graph with the regression line. Note how much better the fit of the regression line with the two outlying cases discarded (R^2 =0.4735).

Identifying and Dealing with Outliers

Apart from visual methods, statistical tools for identifying regression outliers abound. The more commonly known are Mahalanobis distance and Cook's distance. The former measures the distance of a case from the centroid of the remaining cases where centroid is the point created by the means of all variables in a multidimensional space.

$$\text{Mahalonobis distance} = (n - 1)(h_i - 1/n)$$

where n is the number of observations and h_i is the leverage value for ith case derived from the diagonal of the hat matrix $(\mathbf{X'X})^{-1}\mathbf{X'}$.

Cook's distance is another influence measure that reflects the change in the estimates of regression coefficients if the ith case is removed.

$$\text{Cook} = \frac{\text{(h x deleted residual square)}}{\text{(k x residual mean square)}}$$

Figure 3 vividly depicts the outlying observations of the 47th and 48th cases, which Mahalanobis distances are 6.052 and 12.104, respectively, indicating a departure from other cases. Cook's distances for the two cases are 1.039 and 0.664, as compared with the others falling below 0.2.

To circumvent effects of outlying observations, one could remove those cases from the sample, but this sacrifices important information about the outliers.

Table 1. SAT scores and Graduation Rate (GRADRATE)

Case	SAT	GRADRA
1	1152.00	74.40
2	1121.00	69.00
3	1099.00	69.00
4	1069.00	39.00
5	1060.00	68.00
6	1050.00	53.50
7	1044.00	34.00
8	1028.00	41.80
9	1027.00	49.00
10	1026.00	30.00
11	1025.00	47.00
12	1019.00	69.00
13	1009.00	46.00
14	1006.00	50.00
15	1004.00	48.00
16	1000.00	27.00
17	1000.00	45.00
18	998.00	64.00
19	980.00	53.00
20	977.00	34.00
21	968.00	32.00
22	958.00	45.00
23	953.00	46.00
24	927.00	47.00
25	921.00	28.00
26	919.00	44.00
27	918.00	36.00
28	917.00	46.50
29	900.00	50.00
30	892.00	51.00
31	890.00	29.00
32	885.00	25.40
33	876.00	31.00
34	873.00	44.00
35	866.00	41.00
36	857.00	23.00
37	855.00	39.00
38	846.00	37.00
39	831.00	23.00
40	809.00	32.00
41	806.00	12.00
42	799.00	27.00
43	795.00	42.40
44	777.00	41.00
45	760.00	23.00
46	677.00	17.00
47	598.00	72.00
48	464.00	44.10

Deletion of outliers should not be contemplated when the number of cases is substantial. A more positive treatment is to apply Robust Regression techniques that minimize influence of outliers for model estimation.

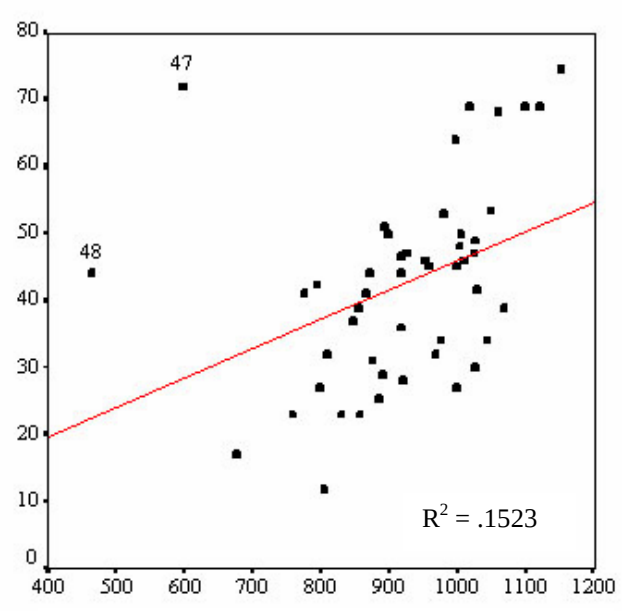

Figure 1. Outliers In: Scattergram of Average SAT and Graduation Rate.

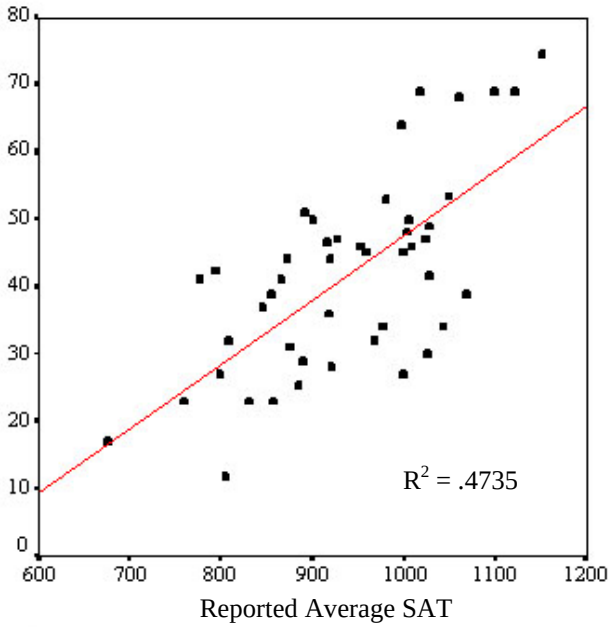

Figure 2. Outliers Out: Scattergram of Average SAT and Graduation Rate.

One of the Robust Regression modeling techniques is based on an MM-estimate computational strategy introduced by Yohai, Stahel and Zamar (1991). The Robust MM Regression method generates highly robust estimates with minimized influence of the outlying cases.

Table 2 lists the model estimates and goodness of fit of the OLS model and Robust MM model using only the SAT score the predict the graduation rate. Notice that the intercept is not statistically significant in the former model. While keeping the two outlying

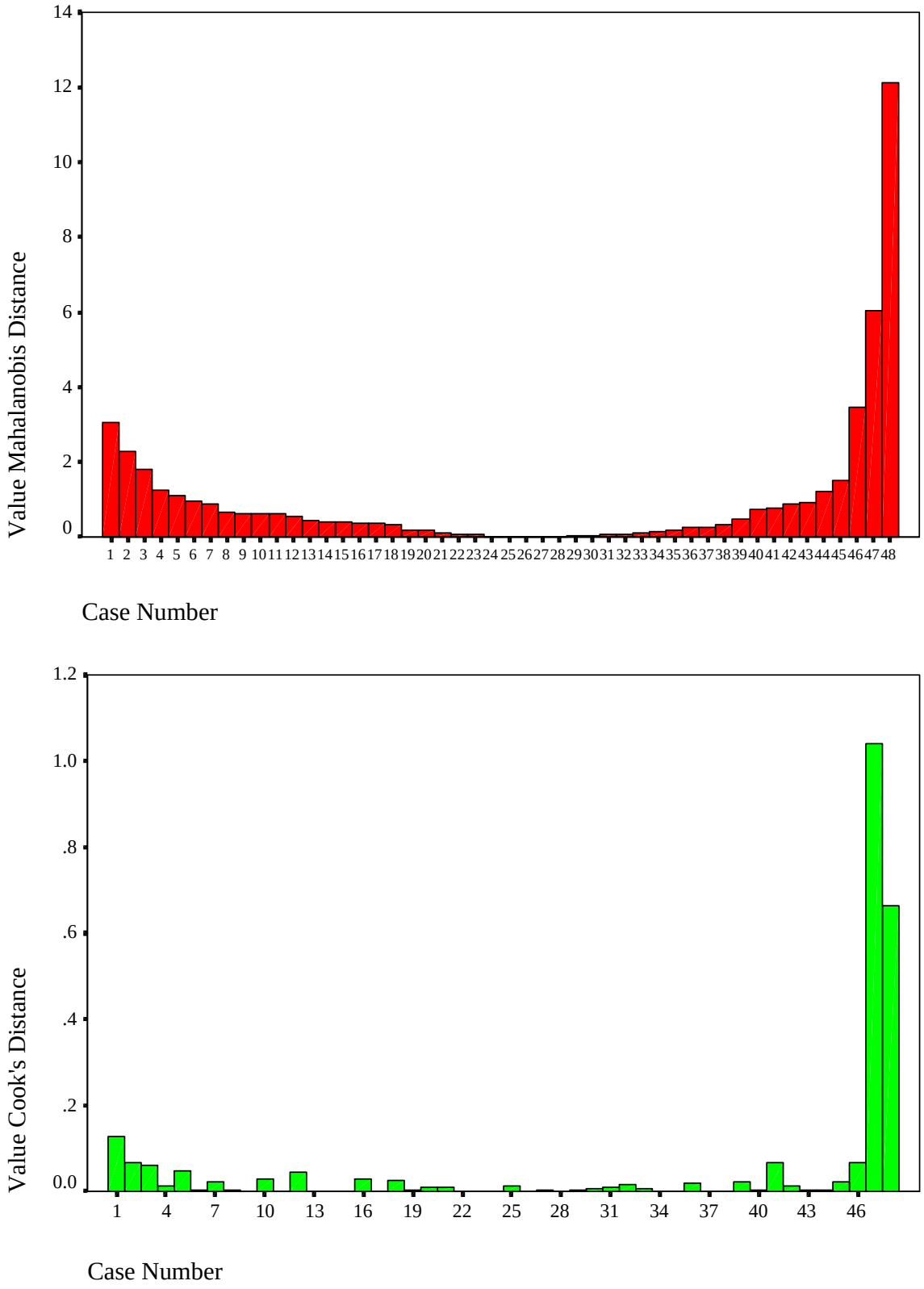

Figure 3. Mahalanobis Distinaces and Cook's Distances.

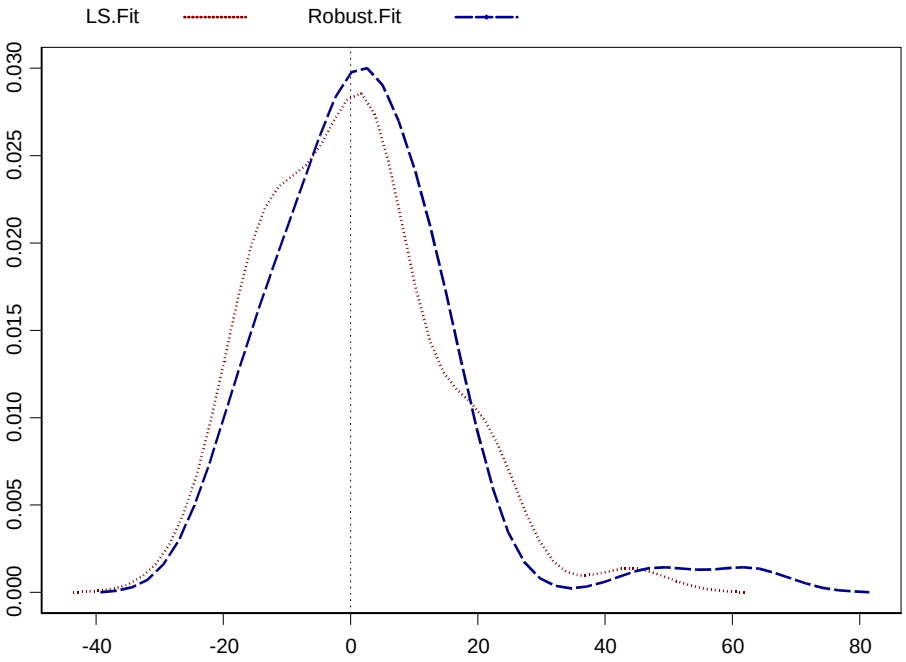

Figure 4. Comparing Densities of Residuals between Robust MM-estimator and Least Square

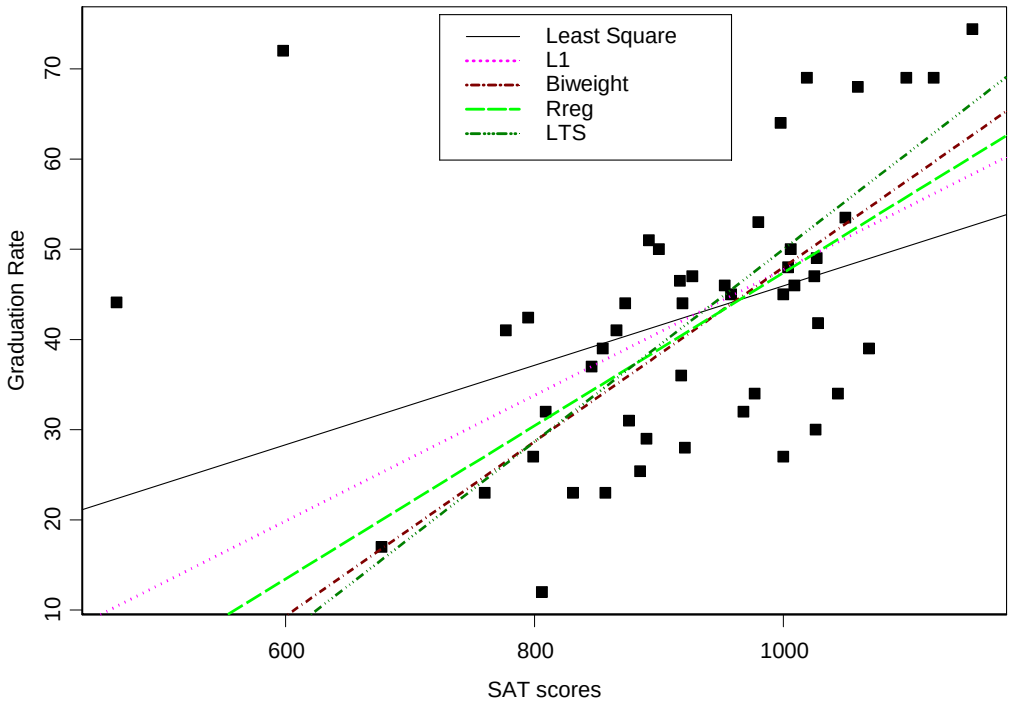

Figure 5. Comparing OLS and Various Robust Estimators

Table 2. Comparison of OLS and Robust Models

OLS	Value	Standard Error	t value	Pr(>|t|)
(Intercept)	1.9170	14.2486	0.135	.893564
SAT	0.0440	0.0153	2.875	.006098
LS Fit :	**0.1523**			
Robust MM				
(Intercept)	-48.2586	16.8244	-2.868	.006209
SAT	0.0960	0.0178	5.382	.000002
LS Fit :	**0.3151**			

cases, namely the 47th and 48th, the Robust MM model does not assume any "manual error" in the data entry but discounts their high influence in modeling the data. The model fit is improved by more than 100 percent.

Figure 4 illustrates how the density of residuals of the robust model is compared to that of the OLS which has bumps on both sides. Comparatively, the robust estimate is well-centered at zero and pushes the outliers farther away to the right.

There are other Robust estimators like Minimum Absolute Residual (L1) Regression, Least Trimmed Squares (LTS), M-estimation(RREG) and Robust Simple Regression by Biweight (Bisquare). Figure 5 demonstrates the relative fit of these robust models compared to OLS. These models can all be implemented using available functions in the S-PLUS 2000 statistical software package (Mathsoft Data Analysis Products Division, 1999).

In conclusion, this article gives a simple illustration of implementing robust models over conventional OLS in the presence of outliers. We demonstrated how outliers can be identified with simple tools and how to deal with data plagued with outlying cases using robust modeling techniques.

References

Hampel, F. Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. 1986. *Robust Statistics: the Approach Based on Influence Functions*. New York: John Wiley & Sons.

Huber, P.J. 1981. *Robust Statistics*. New York: John Wiley & Sons.

Yohai, V., Stahel, W.A. and Zamar, R.H. (1991) A Procedure for Robust Estimation and Inference in Linear Regression, in Stahel, W.A. and Weisberg, S.W., Eds., *Directions in Robust Statistics and Diagnostics, Part II*, Springer-Verlag, New York.

Mathsoft Data Analysis Products Division. 1999. *S-Plus 2000 Guide to Statistics, Volume 1*. Seattle, WA: Mathsoft

(*continued from page 1*)

accomplishment these days. They were blessed with four children, Jeremy, Max II, Miranda, and Johanna, who are as brilliant, creative, and talented as their dad. Max Martin was my dear friend for almost 20 years. I had the utmost respect and admiration for him both professionally and personally. I am a better person for having known him and I will miss him deeply. Max, from all of us, "Well done, my friend. Well done."

Nancy K. Martin
Associate Dean for Undergraduate Studies
College of Education & Human Development
The University of Texas at San Antonio