Journal of Modern Applied Statistical Methods

Volume 6 | Issue 2 Article 11

11-1-2007

Sensitivity Curves for Asymmetric Trimming Hinge Estimators

D.B. Stark
The University of Akron

J.F. Reed III
Lehigh Valley Hospital and Health Services

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 469- 475

1538 – 9472/07/$95.00

Sensitivity Curves for Asymmetric Trimming Hinge Estimators

D. B. Stark
The University of Akron

J. F. Reed III
Lehigh Valley Hospital and Health Services

Robust estimators have been developed and tested for symmetric distributions via simulation studies. The primary objective was to show that they are more efficient than the sample mean when used in conjunction with asymmetric distributions. Little attention has been given to how they perform on data that are from asymmetric distributions, or from distributions that have inherent anomalies (messy data). Thus, the behavior of hinge estimators using sensitivity curve are examined.

Key words: Robust estimators, adaptive estimators, ancillary statistics, selector statistics.

Introduction

In spite of the considerable bad press that the sample mean (the least square estimator of μ) has received, the standard normal theory statistic performs well when real data are nearly normal. However, robust estimators of location have been and continue to be developed and tested for symmetric long and short-tailed distributions by means of extensive simulation studies. The ancestry of these estimators may be traced to the work of Hogg (1967) and the Princeton Robust Study (Andrews, et al., 1972). The objective of these and other studies was to demonstrate that adaptive location estimators would be more efficient than the sample mean. In general, an adaptive procedure may be characterized as an application approach to data analysis rather than a theoretical application.

The objective is to supplement previous simulation studies of robust estimators, hinge estimators, by examining the behavior of these adaptive location estimators using sensitivity curves first developed by the Princeton Robust Study

David B. Stark is Associate Professor of Statistics. His research interests are experimental design, linear models, mathematical statistics, regression, robust statistics, and topology. Email: dstark@uakron.edu James Reed III is Interim Chief of Health Studies and Director of Research at Lehigh Valley Hospital & Health Network. His interests include applied statistical analyses, medical education, and statistical methods in simulation studies. Email him at: James_F.Reed@lvh.com

Selector Statistics and Adaptive Location Estimators

Characteristics such as skewness, tail length, and peakedness describe the distribution characteristics. In defining tail length and skewness, the notation here is from Hogg (1967, 1982). Define L_α = mean of the smallest α n observations and U_α = mean of the largest α n observations. (For instance, if α = 0.05, then $L_{(0.05)}$ is the mean of the smallest [0.05n] observations). Let B = mean of the next largest [0.15n] observations, C = mean of the next largest [0.30n] observations, D = mean of the next largest [0.30n] observations and E = mean of the next largest [0.15n] observations.

Hogg (1967) defined two measures of tail length, Q and Q_1. These two statistics as selector statistics are used to classify symmetric distributions as light-tailed (uniform [0,1]), medium-tailed (normal (0,1)), or heavy-tailed (double exponential). Both Q, $Q = (U_{(0.05)} - L_{(0.05)}) / (U_{(0.50)} - L_{(0.50)})$, and Q_1, $Q_1 = U_{(0.20)} - L_{(0.20)}) / (U_{(0.50)} - L_{(0.50)})$, are location-free and are then uncorrelated with location statistics like the trimmed means. Values of $Q < 2.0$ imply a light-tailed (uniform) distribution, $2.0 \le Q \le 2$ 2.6 implies a medium tailed-distribution (normal), $2.6 < Q \le 3.2$ implies a heavy tailed distribution (double exponential), and a $Q > 3.2$ implies a Cauchy like distribution (very heavy-tailed distribution). When using Q_1 a suggested

classification scheme is: $Q_1 < 1.81$ (light tailed, $1.81 \le Q_1 \le 1.87$ (medium-tailed), and $Q_1 > 1.87$ (heavy-tailed). Hogg (1982) also defined a third measure of tail length H_3 ,where $H_3 = (U_{(0.05)} - L_{(0.05)}) / (E - B)$. $H_3 < 1.26$ is associated with a uniform distribution, $1.26 \le H_3 \le 1.76$ is generally associated with a normal distribution, and $H_3 > 1.76$ is associated with a double exponential distribution.

Hertsgaard (1979) used Q_2, $Q2 = (U_{(0.05)} - T_{25}) / (T_{25} - L_{(0.50)})$, to classify distributions as left skewed ($Q_2 < 0.7$, symmetric ($0.7 \le Q_2 < 1.4$) and right skewed ($Q_2 \ge 1.4$). H_1, $H_1 = (U_{(0.05)} - D) / (C - L_{(0.05)})$, was proposed by Hogg (1982) and was also found to be useful in classifying skewness of a wide variety of distributions. And, Reed and Stark (1996) proposed two quick-and-dirty skewness measures SK_2 , $SK_2 = (X_{(1)} - XMD) / (XMD - X_{(n)})$, and SK_5 , $SK_5 = (X_{(1)} - XM) / (XM - X_{(n)})$. The form of SK_2 and SK_5 are identical to Q_2 and H_1. The advantage of using either the median XMD (Q_2) or mean XM (H_1) lies in the familiarity of these common location estimators. Note: XMD is the median, XM is the arithmetic mean, T_{25} is the [0.25n] trimmed mean (T_α), $X_{(1)}$ and $X_{(n)}$ are the first and last order statistics.

Reed and Stark (1996) proposed a set of asymmetric linear estimators or hinge estimators, defined using the following scheme. Set a total trimming proportion to be trimmed from the sample, α. Determine a proportion to be trimmed from the lower end of the sample (α_l) using the following: $\alpha_l = \alpha [UW_x / (UW_x + LW_x)]$, and the upper trimming proportion, $\alpha_u = \alpha - \alpha_l$., where UW_x and LW_x be the numerator and denominator portions of the selector statistic X and define eight adaptive location estimators as:

Estimator	α	α_l
HQ	0.10	
$\alpha_l = \alpha [UW_Q / (UW_Q + LW_Q)]$		
HQ_1	0.10	
$\alpha_l = \alpha [UW_{Q1} / (UW_{Q1} + LW_{Q1})]$		
HH_3	0.10	
$\alpha_l = \alpha [UW_{H3} / (UW_{H3} + LW_{H3})]$		
HQ_2	0.25	
$\alpha_l = \alpha [UW_{Q2} / (UW_{Q2} + LW_{Q2})]$		
HH_1	0.10	
$\alpha_l = \alpha [UW_{H1} / (UW_{H1} + LW_{H1})]$		
HSK_2	0.10	
$\alpha_l = \alpha [UW_{SK2} / (UW_{SK2} + LW_{SK2})]$		
HSK_5	0.25	
$\alpha_l = \alpha [UW_{SK5} / (UW_{SK5} + LW_{SK5})]$		

In the Princeton Robust Study (Andrews, et al, 1972), sensitivity curves were introduced to provide a basis for comparing estimators. The notion behind a sensitivity curve is to show how the value of a particular estimator is affected by an outlier. The method of construction is fairly straight forward. Start with a symmetric sample that is centered about a given value. In this article, the sample consisted of forty nine points (beginning at -4.8 and ending at 4.8) symmetrical about zero. Then add another point to the sample to see how the value of the estimator is affected. The added point ranged from -9.0 to 9.0. The horizontal axis represents the value of the added point while the vertical axis represents the value of the estimator at that value of the added point.

Results

The sensitivity curve for the sample mean is shown in Figure 1. Note that the curve is a straight line, suggesting that the value of the mean changes linearly with the value of the added point. As the value of the added point increases away from zero, the value of the mean does also. The larger the added value is, the larger the change in the mean value. There is no bound to the influence of the added point.

The Median

The sensitivity curve is given in Figure 2. Note here, that the change in the value of the median is bounded. If the added point is one of the two middle values then it has a direct influence. However, if it is not one of those two values its influence is bounded regardless of the size of the value.

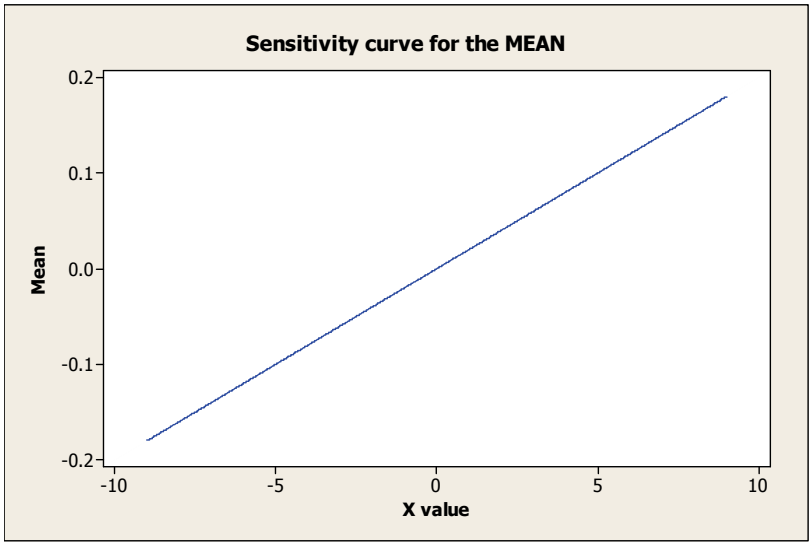

Figure 1.

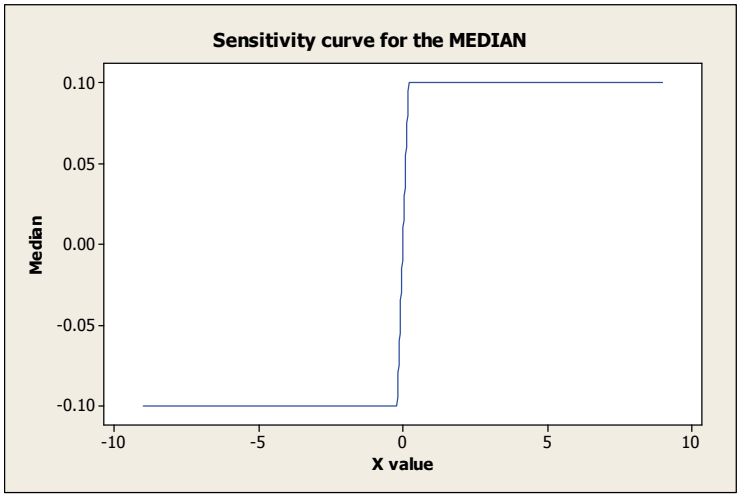

Figure 2.

Next, consider a 15% trimmed mean. The sensitivity curve is shown in Figure 3. As might be expected, the added point has a wider range direct influence. However, once outside of that range of values, the influence of the added point is bounded.

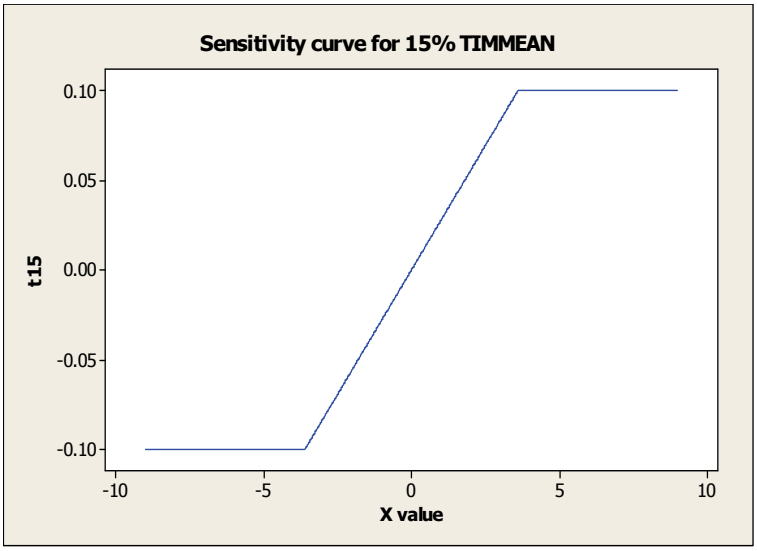

Figure 3.

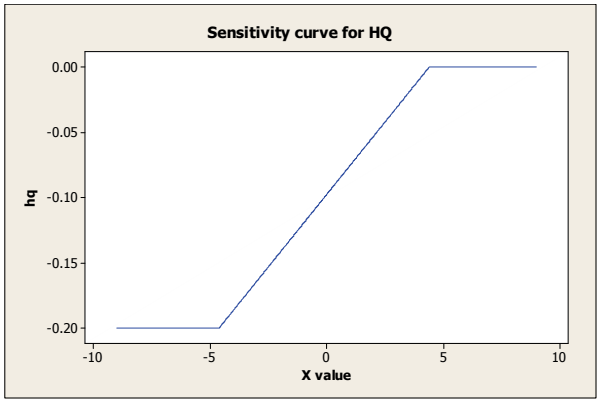

Figure 4.

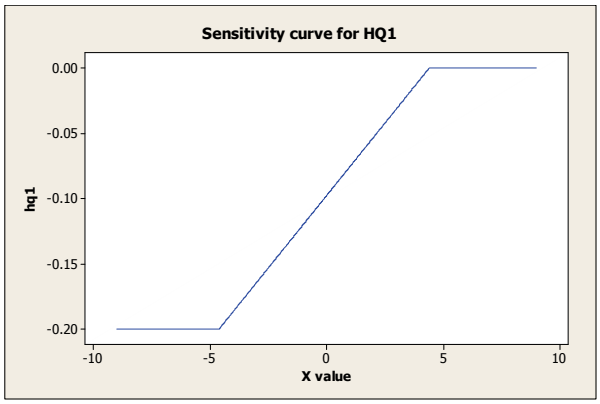

Figure 5.

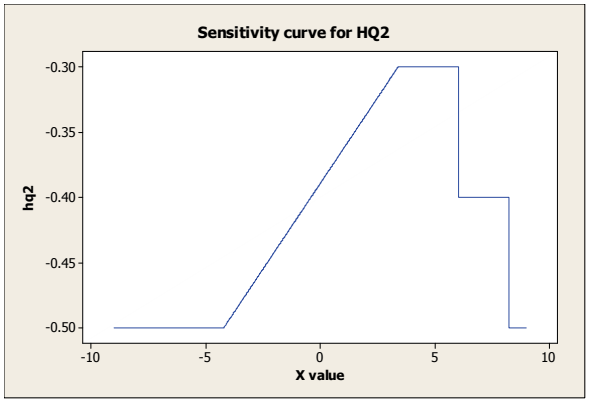

Figure 6.

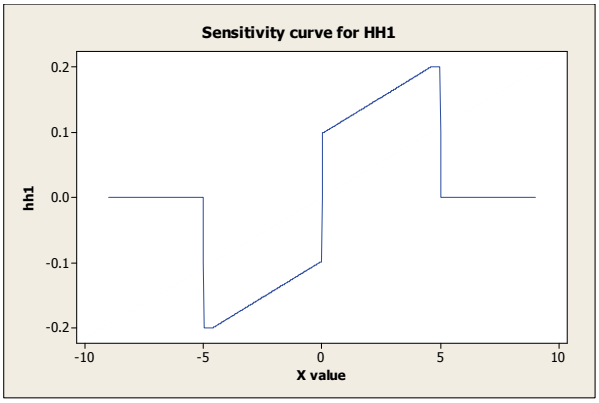

Figure 7.

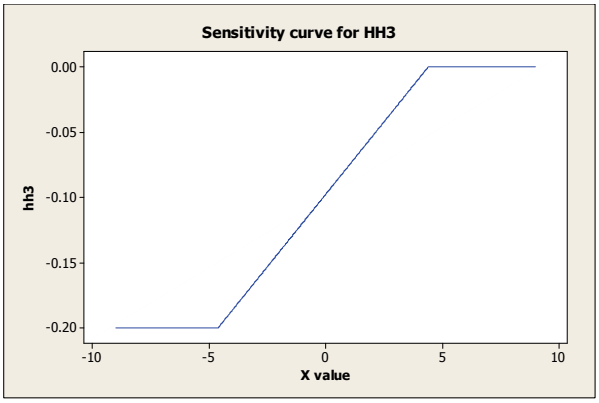

Figure 8.

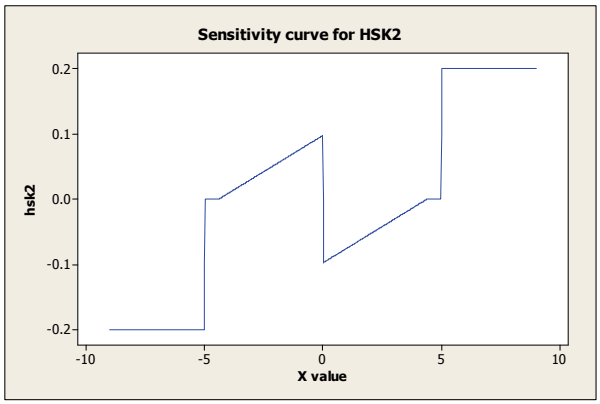

Figure 9.

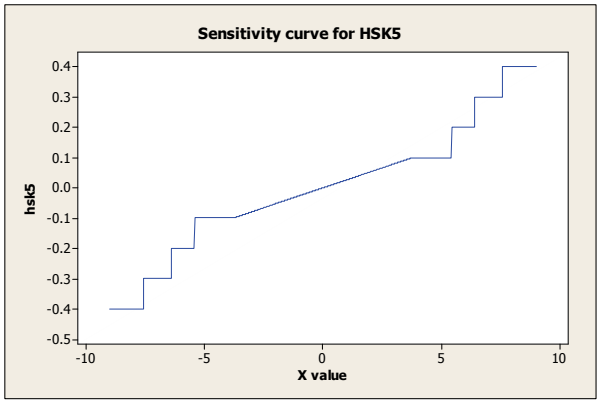

Figure 10.

The sensitivity curves for the seven estimators defined above are given in Figure 4-10. Note the sensitivity curves for HQ, HQ1, HH3 suggest that the adaptive trimming causes the value of the estimator to decrease only. These estimators are not reacting symmetrically to the sample. The estimator HQ2 is just kind of weird. However, HH1, HH3, HSK2 and HSK5 are at least symmetric in their reaction to the value of the added point. The estimator HH1 has a somewhat unique property that when the value of the added point gets a bit outside the symmetric part of the sample, its influence is zero. For data with contaminated with large outliers, this might be a very attractive trait. The estimator HH3 appears to act like a trimmed mean.

Conclusion

Real-world data sets may be described as messy with everything but a normal distribution presenting to the data analyst. From a methodology point rather than a theoretical basis, reasonable alternatives should be available. In the asymmetric data distributions faced on a daily basis, estimators that adapt themselves to the data may be formulated and used. Adaptively trimmed means can correct for uncontrollable data anomalies.

References

Andrews D. F., Bickel P. H, Hampel F. R., Huber P. H., Rogers W. H., and Tukey J. W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press.

Hertsgaard D. M. (1979). Distribution of asymmetric trimmed means. *Comm. Stat-Simul.Comp, 8*, 359-367.

Hogg R. V. (1967). Some observations on robust estimation. *Journal of the American Statistical Association, 62*, 1179-1182.

Hogg R. V. (1982). On adaptive statistical inferences. *Comm. Stat. - Theory Method, 11*, 2531-2542.

Reed J. F. and Stark D. B. (1996). Hinge estimators of location: Robust to asymmetry. *Computer Methods and Programs in Biomedicine*, *49*, 11-17.