Running Head: Robust Statistics

Robust Statistics:

What They Are, Why They Are So Important

Paper to be presented at

Southwest Educational Research Association 32nd Annual Conference, San Antonio, TX

February 4-7, 2009

By

M. Sencer Corlu

Texas A&M University

"Modern" statistics may generate more replicable characterizations of data, because at least in some respects the influences of more extreme and less representative scores are minimized. The present paper explains both trimmed and winsorized statistics, and uses a mini-Monte Carlo demonstration of the robust regression analysis as well as describes other computer-intensive methods such as jackknifing and bootstrapping.

Almost everyone knows about the average or the mean even if they have not heard of any other concept in statistics. Mean is the very first statistical concept taught in schools, and students are told that it is a great tool to represent a whole bunch of numbers. Practically speaking, mean is easy to compute, and its calculation helps students' arithmetic skills improve. However, is it really true that the mean is the greatest tool to represent any set of numbers?

Comparison of mean and median in terms of their robustness:

Turkish people have all the right to be proud of their countryman Hedo, not only because he is one of the very few European players in NBA, but also because of his progress as a basketball player in 2008-2009 season. A skeptical PhD student from Turkey wanted to see if Hedo really deserved the appreciation of 70 million Turkish people. In-between his hectic schedule of statistics courses, he could only find enough time to watch one game per season, and decided to record the number of points Hedo scored against the NBA champion of the previous year. He assumed that would be a good estimate of his performance during that particular year. He formed the following table and started to wait Orlando Magic game against Boston Celtics in 2008-2009 season.

Table 1. Hedo's score summary between 1999-2008

Season	Score against Champ.	Number of Years in NBA
1999-2000	8	1
2000-2001	9	2
2001-2002	10	3
2002-2003	10	4
2003-2004	11	5
2004-2005	12	6
2005-2006	12	7
2006-2007	13	8
2007-2008	14	9

The busy PhD student believing the mean would be a great way of representing Hedo's performance as a basketball player, decided to report only Hedo's statistics which are the moments about the mean.

Table 2. Moments about the mean statistics values for Hedo sample between 1999-2008

Statistic	Value
$\bar{X}$	11
SD_x	1.94
Skewness	0
Kurtosis	-0.8
Pearson r in relation to experience as years	0.99

Over the past 9 years, he prepared different tables with similar numbers filling it. With no knowledge of how well Hedo has been playing so far this season, our skeptical PhD student, with consideration of the Pearson r score being almost equal to 1, expected Hedo to have a score of 15 or so during the game against Celtics. However, Hedo scored just 45 points which will probably lead him to be the second Turkish player in an All-Star game. Our PhD student had to start over from scratch to form a new table of statistics.

Table 3. Moments about the mean statistics values for Hedo sample between 1999-2009

Statistic	Value
$\bar{X}$	14.4
SD_x	10.91
Skewness	3
Kurtosis	9.24
Pearson r in relation to experience as years	0.66

Later on, he wondered on what statistic Hedo's latest performances had the greatest effect. So, he calculated the percent change for $\bar{X}$ and SD_x as 25% and 462%, respectively. Apparently, the latest performance of Hedo had an increasing effect on the increasing moments about the mean. Although our PhD student didn't have the opportunity to test his hypothesis for the third moment about the mean (since *Skewness* in the first data was 0), he extended his percent change calculations to the fourth moment about the mean, and found out that there was a 1569% change in the *Kurtosis* value. He thought perhaps it was time to update his knowledge, since he has been in graduate school for 10 years now, and decided to read more about modern statistical methods.

However, not only modern statistical methods are robust statistics. The sample median could indeed have been a great choice to estimate Hedo's overall performance as an NBA player. The median scores before and after this year's game were 11 and 11.5, respectively. These values were not only very close to each other but also similar to the mean of Hedo's scores before this year's game. Hedo's performance at this year's game was definitely an outlier (Was it really? How would we know this mathematically?), but that had little effect on the sample

median on the contrary to its effect on all the moments about the mean. Two questions emerge at this point:

1) **Why is *median* a robust procedure whereas the mean isn't?**

The answer to the first question is trivial, because when calculating the mean, one has to consider the effects of each and every data point including the outliers, whereas the median disregards the actual data and concerned with the positions, instead.

$$\lim \bar{x} = \lim \frac{x_1+x_2+\cdots x_{n-1}+x_n}{n} \qquad (1)$$

As you see from the formula of the mean, when all but x_n are fixed, and when x_n goes to infinity, the $\bar{x}$ approaches to infinity. Shortly, one bizarre x value can change the sample mean dramatically. One of the many ways to determine whether a statistic is robust or not, is to look at the shape of the influence function to describe the effect of one outlier (Reid, 2006) The influence function is actually the derivative of a functional, and it measures the effect of a small perturbation on a certain statistic. Read Wilcox (2005) for a comparison of the influence functions of some robust statistics. Another method to test for robustness is to simulate the contaminated data sets and see how well the estimator does (Rousseeuw,1991). We will try this with a Monte Carlo simulation in this paper. The easiest method however, is the tool named *breakdown point*. As defined by Donoho & Huber (1983), the breakdown point is the smallest amount of contamination that may cause an estimator to take large bizarre values. The breakdown for the average is 11% (1 in 9) for Hedo example since one outlier was enough for the average to make the nature of the data abnormal. However, the breakdown point for the median is 50%, and that equals to at least five scores (five outliers) that are needed to divert the

value of the median. For infinitely large number of data points, breakdown point for the mean approaches to the lim 1/n =0, whereas the lim (n-1)/2n stays 0.5 for the median.

2) **How can an outlier be detected mathematically?**

An effective method to answer the second question is called jackknife algorithm (Thompson, 2008) In this method, after the initial calculation of the sample statistic, one case in turn is dropped, and the statistic is recalculated for the new subset. Some additional computations provide a confidence interval for the statistic. For example, the table shows the Lower and Upper confidence intervals around the mean calculated by an Excel add-in for Hedo's points as they were given earlier.

Table 4. Jackknife Statistics for Hedo sample between 1999-2008

Jackknife Statistics	
mean	14.40
std. error	3.45
alpha	0.05
t	1.83
LCL	8
UCL	21

Any score not in the range of 8 to 21 is considered as an outlier for Hedo case.

Robust statistics on multivariate statistics:

Let's continue our discussion on robust statistics with multivariate statistics. The following scatter graphs show the linear relationship between the number of years of experience of Hedo in

NBA, and his performance in terms of baskets he scored at a certain game. It is clear from the figure 1 and figure 2 that the outlier (score in 2008-2009 was 45 points) greatly affected the line of best fit. Comparing r^2 values of both sets tells us that there was approximately 50% decrease in the linearity ($0.99^2 - 0.66^2 = 0.55$). We are now done with Hedo for now.

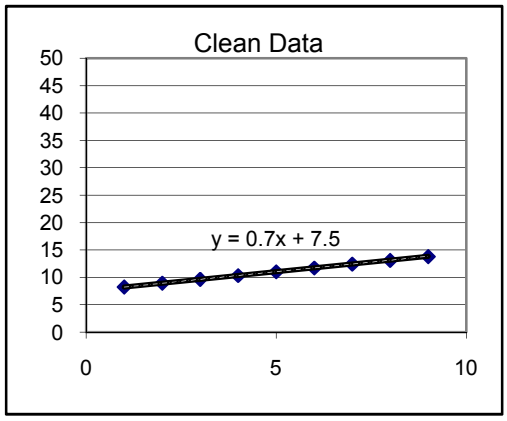

Figure 1: Hedo Scattergram Case 1

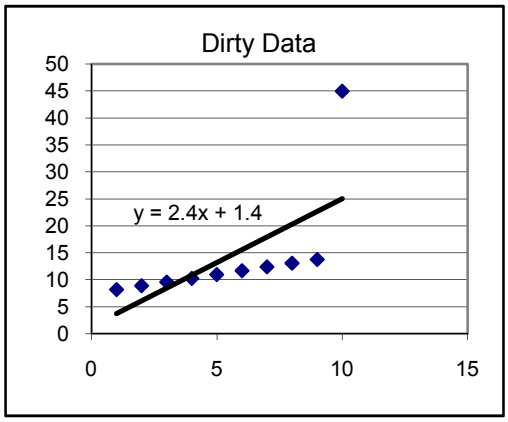

Figure 2: Hedo Scattergram Case 2

Is there any other regression method that is more robust than the method of least squares (LS)?

Rousseeuw (1984) suggests minimizing the median of squared residuals (LMS) instead of trying to minimize the sum of squared residuals would produce more robust estimation. After all, as we

explained earlier the median is more robust to change compared to the mean, thus compared to the sum. Given the data generating process $Y_i = \beta_0 + \boldsymbol{\beta_1} X_i + \varepsilon_i$, a 10,000 times iterated Monte Carlo simulation of the slopes of some other 11 samples with parameter values β_0=1 (y-intercept), $\boldsymbol{\beta_1 = 5}$ ($\boldsymbol{slope}$), and α=10 and an outlier factor of 100 added to the 11th sample produced the following two figures for LS, and LMS, respectively.

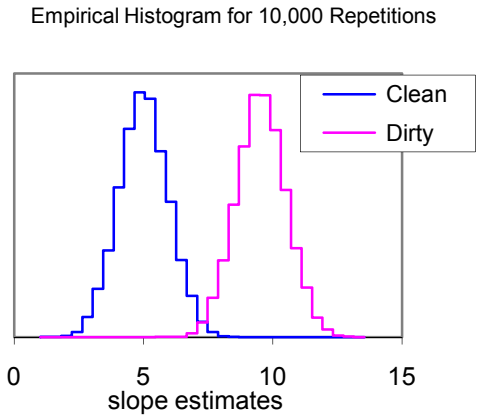

Figure 3: Least squares regression

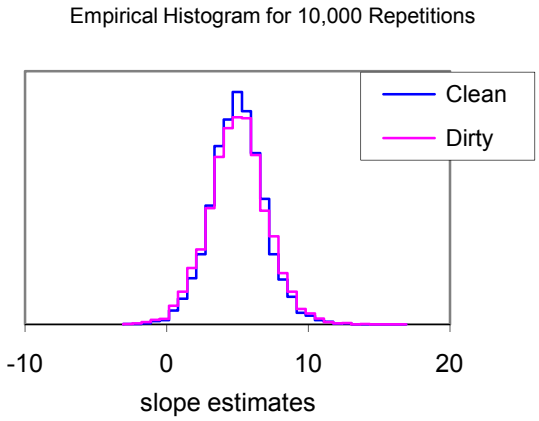

Figure 4: Least Median of Squares

However, although LMS estimates the slopes better, it is way less precise than the LS. The following two figures show this fact with a Monte Carlo simulation of slopes for clean and contaminated data with the same parameters, i.e., $\boldsymbol{\beta_1 = 5}$.

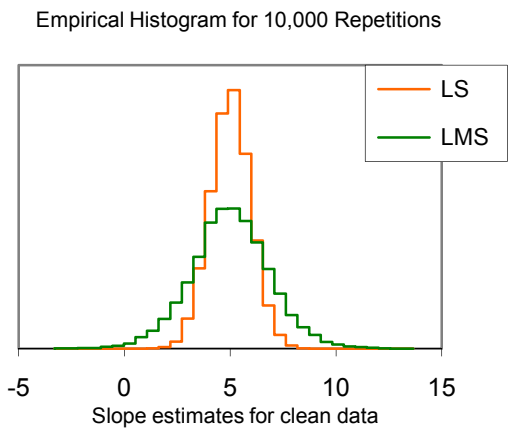

Figure 5. Comparison of LS and LMS on clean data

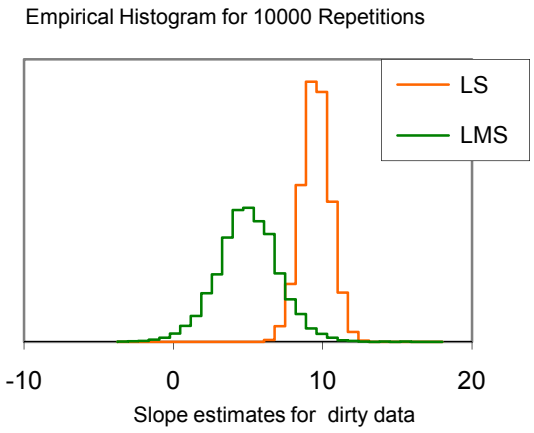

Figure 6. Comparison of LS and LMS on clean data

The Monte Carlo simulations reveal the fact that ordinary least squares is not a robust estimator, that it is weak against extreme values of data. On the other hand, LMS is not much affected by the outliers; however it suffers from low precision.

Couldn't we just get rid of the outlier data instead going through the complicated process of LMS? In fact, another robust estimator called Least Trimmed Squares (LTS) follow this method. Instead of working on LTS, we will leave this to the reader, and go back to univariate data to talk more about trimming.

Trimming and Winsorizing

Hedo is not the only Turkish basketball player in NBA, nor is he the first player to play an All-Star game. Mehmet Okur, aka Memo has been selected as an All-Star player in 2007-2008. A random sample of the points he scored in 10 games throughout his career in NBA are 22, 22, 28, 27, 19, 2, 19, 88, 100, and 1. Some of the scores are more likely than the others to be selected from a pre-assumed normal distribution of Memo's scores and some of these scores appear to be so bizarre that they do not belong to the Memo's score population with a normal distribution. We speculate the reasons for the bizarreness as

1) Inaccurate data entry, because nba.com is maintained by a careless webmaster
2) After looking at the details of the game, we figured out that Memo was injured and left some games early.

Otherwise, without being able to explain the reasons for the bizarre data, we should merely conclude that Memo performed within his potential, and should keep the data as they are. This leads us to treat the outliers as other data, and not discriminate against them.

After finding a good valid reason to treat outliers as data that contaminated Memo's overall performance, we sort the data in an increasing fashion, and trim the 20% of the data from both ends. This leaves us only 60% of the original sample, which includes 19, 19, 22, 22, 27, and 28

as our new trimmed sample. Wilcox (1997) suggests 20% as the best amount of trimming for general purposes. Considering the reasons behind the bizarreness of our data, we keep his advice in this case. The main advantage of using a trimmed mean is that we do not risk the normality presumption, because if Memo's scores throughout his entire career are normally distributed, the trimmed mean will be equal to the mean. (Erceg-Hun & Mirosevich, 2008)

What happens if we decide to trim 50% of the odd-numbered sample? We are then left with the median only, as our trimmed mean. The other extreme trimming percent merely gives us the sample mean with 0% trimming.

Alternatively, we can replace the trimmed samples with the first and last samples in our trimmed sample. Memo's winsorized sample becomes **19, 19,** 19, 19, 22, 22, 27, 28, **28, and 28**. By doing so, we are paying more attention to those scores near the centre by giving less importance to the extremes. The following table compares some statistics of our original sample, trimmed sample, and winsorized sample.

Table 5. Comparison of robustness methods in moments about the mean statistics for Memo's scores

Original Sample		20% Trimmed Sample		20% Winsorized Sample	
Statistic	Value	Statistic	Value	Statistic	Value
$\bar{X}$	32.8	$\bar{X}$	22.83	$\bar{X}$	22.5
Median	22	Median	22	Median	22
SD_x	33.65	SD_x	3.87	SD_x	4.18
Skewness	1.47	Skewness	0.5	Skewness	0.26
Kurtosis	1.05	Kurtosis	-1.73	Kurtosis	-2.12
Range	99	Range	9	Range	9

In the case of choosing our trim percentage to 25%, the range automatically becomes IQR (inter-quartile range), which is the difference between 75^{th} and 25^{th} percentiles of the data set. This is given as a robust dispersion descriptive statistics in Thompson (2008).

Bootstrapping

Bootstrapping is a resampling method performed by the help of a computer software, such as an Excel add-in. Consider Memo's sample of scores: 22, 22, 28, 27, 19, 2, 19, 88, 100, and 1 with a sample mean of 32.8. An Excel add-in replaces the original sample with a randomly selected 12 observations. One random sample generated by bootstrapping might be 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1 with a mean of 1 or another might be 100 repeated 12 times. This process of generating bootstrap samples from the original sample is repeated thousands of times, to get a better approximation of the sampling distribution of the mean (or any other statistic). With bootstrapping method p-value calculations, hypothesis, or even forming confidence intervals around the mean or the effect sizes become more accurate and replicable even if the assumptions around the population is violated (Erceg-Hun & Mirosevich, 2008). Bootstrapping is used to get a more robust estimation of sampling distribution than the theoretical distributions. This is most helpful when the population is not assumed to be normal or homogenous in variance.

Erceg-Hun & Mirosevich (2008) reports the normality assumption is rarely met in practice, supporting Micceri (1989) who found out that real data are more likely to resemble an exponential curve rather than a normal. This causes many studies wrongly rejecting the null hypothesis or reporting low power values. Besides normality, there is another issue which is pretty important even if the population is rightly assumed to be normal. It is the existence of differences between random variables. Wilcox (1998) clearly mentions that a small diversion from the homogeneity of variance principle may cause a dramatic low power value, even if the sample sizes are equal. He rightly asks the question, "How many discoveries have been lost by

ignoring modern statistical methods?" (Wilcox, 1998, p.1) and states that many nonsignificant results have been found to be significant because modern statistical methods were ignored.

Computer Software

R is a statistical software package, which is free and able to do all the procedures described in this paper. ZumaStat claims to be more user-friendly than R, and uses SPSS and Excel for robust analysis. Other than these two, several excel add-in's are available on the Internet, and are available through my webpage at http://people.tamu.edu/~sencer, where a small description and assessment in terms of their ease of use is included.

Conclusion

As Thompson (2008) warns, outliers are not merely evil people who distort all statistics for all variables. In some cases, outliers and their effects are more important than the rest of the data, and the mean would a better estimate regarding the purpose of our research (e.g. the women who were having a healthy life with HIV virus). Osborne, Jason & Overbay (2004) say an unlikely case might shed light on an important principle or issue, which might be too valuable to ignore. In cases where outliers contaminate the general nature of the data set, robust statistical methods are considered as modern and generate more replicable results than classical methods. Given all the recent developments in computer technology, it is easier than ever to use robust methods, and perhaps it is time for all the researchers to consider effects of the modern methods.

References

Donoho, D. L., & Huber, P. J. (1983). The notion of breakdown point. In P. Bickel, K. Doksum, & J. L. Hodges Jr (Eds.), *A Festschrift for Erich Lehmann* (pp. 157-184). Belmont, CA: Wadsworth.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods. *American Psychologist*, *63*, 591-601.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological bulletin*, *105*, 156-166.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, *9*(6). Retrieved from http://PAREonline.net/getvn.asp?v=9&n=6

Reid, N. (2006). Influence functions. In (No Ed.), *Encyclopedia of statistical sciences*. Retrieved February 1, 2009, from http://mrw.interscience.wiley.com/emrw/9780471667193/ess/article/ess1240/current/pdf

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of American Statistical Association*, *79*, 871-880.

Rousseeuw, P. J. (1991). Tutorial to robust statistics. *Journal of Chemometrics*, *5*(1), 1-20.

Thompson, B. (2008). *Foundations of Behavioral Statistics: An Insight-Based Approach* (Paperback ed.). NY: Guilford.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods?. *American Psychologist*, *53*, 300-314.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego:CA: Academic Press.