UNDERSTANDING LEAST ABSOLUTE VALUE IN REGRESSION-BASED DATA MINING

Matt Wimble[1], Michele Yoder[2] and Young K. Ro[3*]

[1]College of Business, University of Michigan - Dearborn (313) 583-5286
[2]College of Business, University of Michigan - Dearborn (313) 583-6316
[3]College of Business, University of Michigan - Dearborn (313) 593-4078

ABSTRACT

This article advances our understanding of regression-based data mining by comparing the utility of Least Absolute Value (LAV) and Least Squares (LS) regression methods. Using demographic variables from U.S. state-wide data, we fit variable regression models to dependent variables of varying distributions using both LS and LAV. Forecasts generated from the resulting equations are used to compare the performance of the regression methods under different dependent variable distribution conditions. Initial findings indicate LAV procedures better forecast in data mining applications when the dependent variable is non-normal. Our results differ from those found in prior research using simulated data.

KEYWORDS

L1-Norm estimation, least absolute value, variable selection, data mining, robust regression

1. INTRODUCTION

"Data mining is a blend of concepts and algorithms from machine learning, statistics, artificial intelligence, and data management. With the emergence of data mining, researchers and practitioners began applying this technology on data from different areas such as banking, finance, retail, marketing, insurance, fraud detection, science, engineering, etc., to discover any hidden relationships or patterns." [14; p.969]. Optimizing the selection of variables in a regression model has long been a subject of interest for empirical scholars [16]. Regression-based variable selection methods such as Stepwise regression [10] were among the first techniques that could be considered "data mining." Applying regression-based variable selection has proven useful in identifying input variables for other techniques such as Neural Network Classifiers [25]. Problems that arise out of violations of regression assumptions have long been a subject of interest [21], ever since computerization facilitated automated variable selection techniques.

Least squares (LS) regression estimates have been widely shown to provide the best estimates when the error term is normally distributed (e.g. [18], [19], [27]). However, instances of violations of the underlying normality assumption have been shown to be quite common. In both finance and economics, the existence of non-normal error terms has been shown to exist [23]. Investment returns have been known for some time to violate assumptions of normality [9], [24], [30]. Non-normality exists in biological laboratory data [15], psychological data [22], quality control applications [33], and RNA concentrations in medical data [8]. Statistics textbooks

DOI : 10.5121/ijdkp.2016.6301

written for applied researchers claim normality assumptions are adequate [31], despite the well-known problems with these assumptions. Established financial theory includes normality violations in option pricing models where lognormality is expressed as a model assumption [1]. Natural phenomena such as tornado damage swaths, flood damage magnitude, and earthquake magnitude have been shown to exhibit normality violations [29].

LS parameters are calculated by minimizing the sum of the squares of the distance between the observed and forecasted values. Least absolute value (LAV) parameters are calculated by minimizing the absolute distance between observed and forecasted values. Although LAV was proposed [3] earlier than LS [20], LS has been adopted as the most widely used regression methodology. Charnes, Cooper, and Ferguson [4] are given credit for first utilizing the simplex method to solve a LAV regression problem. In fact, most scholarly work on approaches to solving the LAV regression problem included variations of the simplex method up until the 1990s. The absolute value function used in LAV is a function where the first derivative is discontinuous, precluding the use of calculus to find a general solution. The lack of a general solution for LAV makes the method difficult to study from a theoretical standpoint and study is often limited to simulation methods such as Monte Carlo.

Several simulation studies comparing the performances of LS and LAV in small samples have been done. The studies of Blattberg and Sargent [2], Wilson [32], Pfaffenberger and Dinkel [26], and Dielman [5] have suggested that LAV estimators are 80 percent as efficient as LS estimators when error terms are normally distributed. When the error distributions contain large tails, large gains in efficiency occur with LAV [5]. Stepwise variable selection methods are by far the most common data mining techniques, and hypothesis testing measures for LAV estimates exist to facilitate this procedure [6].

To our knowledge, there is a dearth of scholarly works exploring the application of LAV to the data mining context. This article extends our understanding of the application of LAV to data mining by utilizing observed U.S. demographic data, which has the potential to violate normality assumptions, to compare the accuracy, consistency, and efficiency of LS and LAV estimators in an actual data mining application. We focus on enumerating all regression estimates in order to provide analysis for those using meta-heuristic optimization methods, such as Tabu Search [11]–[13] and Genetic Algorithms [17], to search the solution space.

2. METHODOLOGY

The study was conducted using U.S. State [28] data obtained via Visual Statistics2 supplementary data sets [7]. Variable selection is a combinatorial problem, and for the sake of this study, four different variables were selected, using a genetic algorithm, out of 20 possible variables. This yielded 4,845 possible combinations (size was limited to enable timely enumeration). It is not uncommon for a researcher using this technique to find new insights to try models with upwards of 150 variables. The original dataset contained 132 variables, which were trimmed to keep roughly an even distribution of demographic, economic, environmental, education, health, social, and transportation variables. Criminal, political, and geographic variables were omitted due to the size constraint. The independent variables used are shown in Table 1.

Table 1. Explanatory Factors Used

AvBen	Average weekly state unemployment benefit in dollars
EarnHour	Average hourly earnings of mfg production workers
HomeOwn%	Proportion owner households to total occupied households
Income	Personal income per capita in current dollars
Poverty	Percentage below the poverty level
Unem	Unemployment rate, civilian labor force
ColGrad%	Percent college graduates in population age 25 and over
Dropout	Public high school dropout rate
GradRate	Public high school graduation rate
SATQ	Average SAT quantitative test score
Hazard	Number of hazardous waste sites on Superfund list
UninsChild	Percentage of children without health insurance
UninsTotal	Percentage of people without health insurance
Urban	Percent of population living in urban areas
DriverMale%	Percent of licensed drivers who are male
Helmet	1 if state had a motorcycle helmet law, 0 otherwise
MilesPop	Annual vehicle miles per capita
AgeMedian	Median age of population
PopChg%	Percent population change
PopDen	Population density in persons per square mile

One dependent variable was chosen that was approximately normally distributed. The remaining three dependent variables were randomly chosen from the variables in the original dataset that were not used as independent variables. Distributions were measured using BestFit. Distribution fit is calculated using Chi-Square Test, Anderson-Darling Statistic (A-D), and the Kolmogorov-Smirnov Test (KS). The normal variable used was "average daily hospital cost in dollars per patient" (Hospital Cost). This variable was chosen because the normal distribution represented the best fit in two out of the three tests. The other dependent variables used were "1997 federal grants per capita for highway trust fund and FTA" (Federal Grants), "1996 hospital beds per thousand population" (Hospital Beds), and "1996 DoD total contract awards in millions of dollars" (Defense Contracts). The histograms for the dependent variables are shown in Figures 1-4.

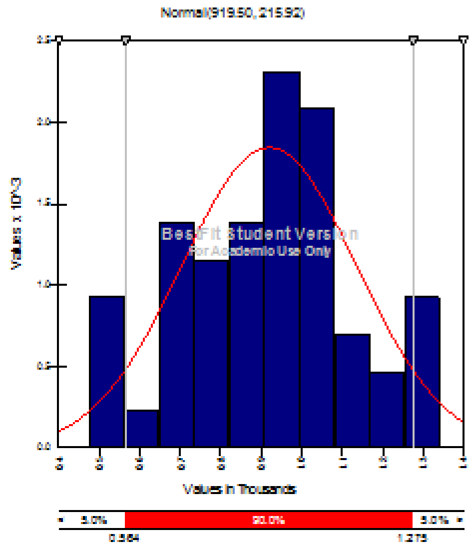

Figure 1. Hospital Cost Histogram

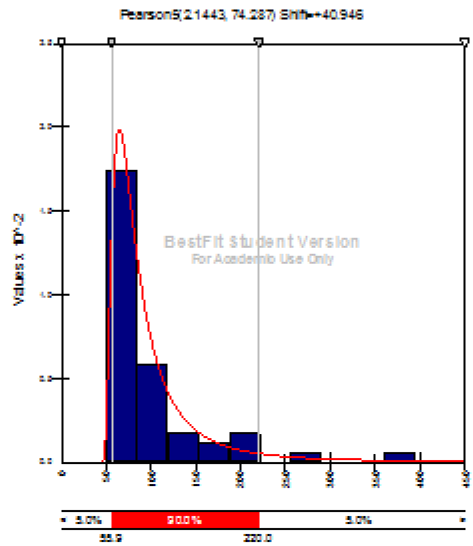

Figure 2. Federal Grants Histogram

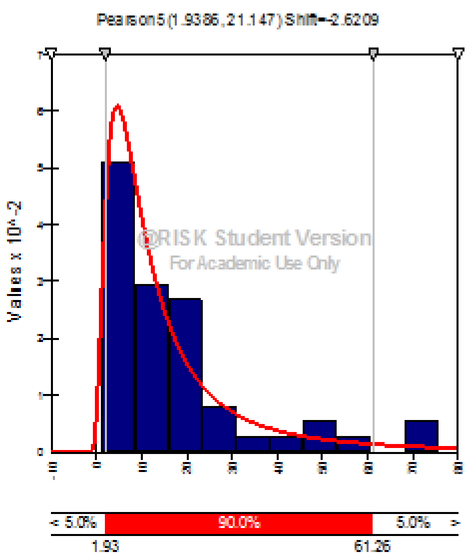

Figure 3. Hospital Beds Histogram

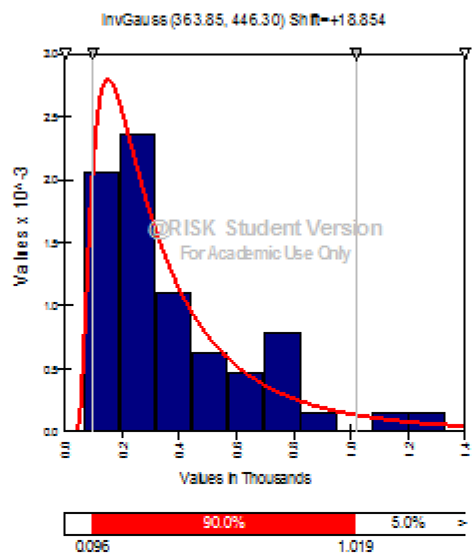

Figure 4. Defense Contracts Histogram

Once the independent and dependent variables were selected, a complete enumeration of all LAV and LS models was performed. Function minimization was performed using the Premium Solver Add-In for Excel by Frontline Systems. Initial models were verified using conventional regression methods to verify validity-of-technique. Data was bifurcated into even 25-state groups, one for training and one for validation.

Performance was measured based on ability to forecast on the validation set values and the model was fit using the other 25 state set. We compared the performance of LS and LAV by assessing relative accuracy, efficiency, and consistency. Accuracy was measured by both mean absolute deviation (MAD) and percentage of LAV forecasts that were closer (%LAV Closer). MAD is defined as:

$$MAD = \frac{\sum_{i=1}^{n} \left| \hat{y}_i - y_i \right|}{n}$$

MAD is a measure of how far the forecasts deviate from the observed values, with a smaller number indicating greater accuracy. Percentage of LAV forecasts that were closer is defined by the number of LAV forecasts that were closer to the true value than the LS forecasts, divided by the number of forecasts (in other words, how often LAV produced a better forecast than LS). Relative efficiency of LAV is defined as:

$RE = \dfrac{RMFE_{LAV}}{RMSFE_{OLS}}$, where $RMSFE_1 = \left(\dfrac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n} \right)^{\frac{1}{2}}$, and where n = number of forecasts.

A number under 100% indicates that LAV is less efficient than LS, whereas a number greater than 100% indicates that LAV is more efficient than LS. Finally, the standard deviation of absolute forecast errors was used to assess consistency.

3. RESULTS

A total of 4,845 regression models were run for each of the dependant variables. The top 1, 2, and 5 percent of the models on the validation sets for both LS and LAV were compared to each other. Specifically, the LAV forecasts with the lowest absolute fitted error were compared with the LS forecasts with the lowest squared fitted error.

Figure 5 displays the Percent LAV Closer and Relative Efficiency results for the top 2% of observations for each of the four dependent variables measured.

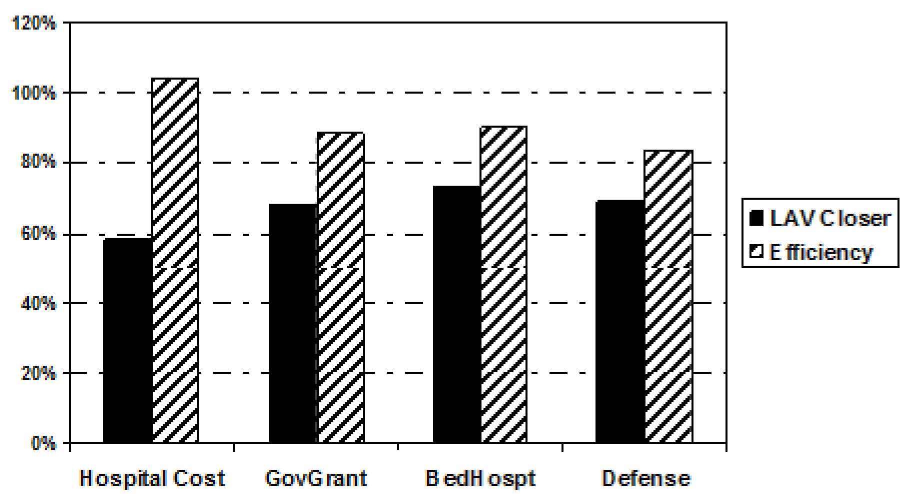

Figure 5. Comparison by Variable for Top 2% of Fits

For every dependent variable except Hospital Cost, LAV produces more accurate results than LS in over 60% of the cases, based on the Percent LAV Closer measures. However, LAV does not produce a more efficient estimator, except in the case of the Hospital Cost dependent variable.

Tables 2 through 5 provide all the accuracy, efficiency, and consistency data for each dependent variable.

Table 2. Comparison of LAV and LS – Hospital Cost

		MAD	%LAV closer	Efficiency	Std. Dev
Top 1% (48 obs)	**LAV**	4077.0	58.3%	96.9%	390.7
	LS	4157.5			147.1
Top 2% (97 obs)	**LAV**	4133.5	46.4%	104.0%	419.7
	LS	4062.6			293.0
Top 5% (242 obs)	**LAV**	4188.1	47.9%	110.4%	611.3
	LS	4001.3			458.5

Table 3. Comparison of LAV and LS – Government Grants

		MAD	%LAV closer	Efficiency	Std. Dev
Top 1% (48 obs)	**LAV**	1179.4	75.0%	85.6%	145.2
	LS	1279.7			103.5
Top 2% (97 obs)	**LAV**	1147.3	68.0%	88.4%	153.7
	LS	1223.8			131.1
Top 5% (242 obs)	**LAV**	1071.2	72.3%	86.6%	141.8
	LS	1153.6			133.7

Table 4. Comparison of LAV and LS – Hospital Beds

		MAD	%LAV closer	Efficiency	Std. Dev
Top 1% (48 obs)	**LAV**	321.9	79.2%	92.0%	20.0
	LS	335.7			19.5
Top 2% (97 obs)	**LAV**	320.7	73.2%	90.0%	23.0
	LS	338.4			17.5
Top 5% (242 obs)	**LAV**	322.9	76.4%	91.0%	21.6
	LS	339.1			12.6

Table 5. Comparison of LAV and LS - Defense Contracts

		MAD	%LAV closer	Efficiency	Std. Dev
Top 1% (48 obs)	**LAV**	5762.0	66.7%	94.0%	1195.3
	LS	6057.0			371.3
Top 2% (97 obs)	**LAV**	5395.5	69.1%	83.6%	1302.6
	LS	6050.7			490.6
Top 5% (242 obs)	**LAV**	5269.3	73.1%	79.7%	1170.5
	LS	6023.4			509.6

Recall that the Hospital Cost dependent variable was selected because it was normally distributed. The remaining dependent variables are markedly non-normal. The MAD and Percent LAV Closer results indicate that LAV outperforms LS in accuracy for all the non-normal dependent variables (Tables 3-5). The MAD values are larger for LS than for LAV, and the Percent LAV Closer is over 66% in every case. For the Hospital Cost variable, LS is more accurate than LAV for both the top 2% and top 5% of observations. Therefore, we can conclude that LAV is more accurate than LS when the dependent variables are not normally distributed.

The relative efficiency values ranged from 79.7% to 110.4%, with the highest values found for the normally-distributed Hospital Costs dependent variable. Therefore, we can conclude that LS is more efficient than LAV when the dependent variables are not normally distributed.

4. DISCUSSION

We found that, when non-normal data are used, LAV is more accurate, less efficient, and more consistent than LS. It is worth noting that comparing performance in this study to Dielman's [5] simulation results becomes problematic in that Dielman used symmetric distributions (i.e. normal, contaminated normal, Cauchy, and Laplace), where the non-normal data in our study exhibited considerable skewness. However, the use of real, skewed data represents a strength of this study; Dielman's study used simulated data rather than real data, which is inherently less normal. This matches the basic pattern found in prior simulation studies with some important differences in the magnitude of the findings.

Specifically, LAV performed at or better than what simulation results tended to suggest in terms of accuracy, with forecasts being closer about 10% more often than in Dielman's [5] study. In relative efficiency terms, LAV performed worse than the simulation would have suggested. Dielman's study showed LAV to have relative efficiency measures in the range of 125%. In this study LS performed only slightly better in relative efficiency terms, with relative efficiency measures ranging from 79-110%. Variation from normality most likely explains differences in relative efficiency from Dielman's findings. With regards to consistency, an interesting finding was that LS produced a more consistent forecast than LAV for all dependant variables used.

Clearly, LAV represents a tradeoff in data mining applications. The results of this study suggest that LAV outperforms LS in terms of accuracy and consistency when data are non-normal. In particular, the data we used are markedly skewed, a not uncommon occurrence in real data. However, LS remains generally more efficient. The degree to which this is an issue depends on the number of observations available. In small data sets, the lack of efficiency is troubling. However, as modern technology continues to advance and allow for the collection of large data sets with numerous observations, the accuracy and consistency of LAV may well outweigh the inefficiency.

5. LIMITATIONS & FUTURE RESEARCH

This paper is not without limitations. The study provides a starting point for using LAV, a computationally expensive procedure, within a regression-based variable section data mining model. The computational requirements are more difficult in the case of LAV since it is not a transform as in the case of LS. Computing LAV is a simplex operation and requires more computational steps than LS. When put in the context of a variable selection model, it becomes computationally more difficult. This shows that performance differences still apply within a variable selection context. However, variable selection itself is computationally expensive. Implied in this context is that given a fixed amount of time, at scale, some variable selection computing must be traded off for estimation computing. A key limitation is that these results do not address how this tradeoff should be managed. Second, inferring the estimator properties with LAV requires the use of Monte-Carlo simulation. Finally, while this study uses real-world data with outcomes from varying distributions, these results are less generalizable than a systematic simulation.

Initial results from this study suggest that further investigation of LAV estimation and other robust regression techniques, within the context of variable selection, are worth scholarly pursuit. A larger sample of forecast tests, from various areas of study (such as biology, manufacturing, medicine, epidimeology, etc.) and reflecting additional violations of normality, are necessary to provide sufficient justification for wide-ranging use of LAV-based regression to select variables. In particular, additional research should be undertaken to determine if the findings of forecast consistency remain and if performance under skewness also remains. The interaction between suboptimal LAV regression estimates within a variable selection metaheuristic, such as Tabu Search or Genetic Algorithms, has not been explored. Additionally, LAV is one of many types of robust regression techniques that could be explored within the same problem framework.

6. CONCLUSION

Given how often normality violations occur in real data, the use of robust estimation techniques such as LAV would seem to be useful in regression-based data mining. Preliminary results suggest that LAV could be useful in regression-based data mining models, but more data is needed to derive substantial conclusions. Simulation studies of this technique are difficult to conduct due to the factorial nature of the number of possible models that need to be controlled. As a result of these difficulties, studies with real data present a promising way to study LAV and other robust techniques as well. Another open question is whether the potential benefits of LAV outweigh the computational overhead of Simplex, versus the guaranteed O(n) of LS, when used within the metaheuristic necessitated by the problem scale.

REFERENCES

[1] Black, P. and Scholes, M., (1973), “The Pricing of Options and Corporate Liabilities”, Journal of Political Economy, 81, 637-654.

[2] Blattberg, R. and Sargent, T., (1971), “Regression with non-Gaussian stable disturbances: some sampling results,” Econmetrica, 39:501-510

[3] Boscovich, R., (1757),”De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operas, ac habentur plura ejus ex exemplaria etiam sensorum impressa,” Bononiensi Scientiarum et Artum Instituto Atque Academia Commentarii, 4:353-396

[4] Charnes, A., Cooper, W. W., and Ferguson, R., (1955), “Optimal estimation of executive compensation by linear programming,” Management Science, 1:138-151.

[5] Dielman, Terry E., (1986),”A Comparison of Forecasts from Least Absolute Value and Least Squares Regression,” Journal of Forecasting, 5:189-195.

[6] Dielman, T. E. and Rose, E. L., (2002), “Bootstrap Versus Traditional Hypothesis Testing Procedures for Coefficients in Least Absolute Value Regression”, Journal Statistical Computation and Simulation, 72:665-675.

[7] Doane, D., Tracy, R, and Matheison, K. (2001), Visual Statistics 2, The McGraw-Hill Companies, Inc.

[8] Dyer, J., Pilcher, C., Shepard,R., Schock,, J., Eron, J., and Fiscus, S, (1999), “Comparison of NucliSens and Roche Monitor Assays for Quantitation of Levels of Human Immunodeficiency Virus Type 1 RNA in Plasma”, Journal of Clinical Microbiology, v. 37, No. 2, pp 447-449

[9] Fama, E., (1965),”The behavior of stock market prices”, Business, 38:34-105.

[10] Efroymson,M. A. (1960) "Multiple regression analysis," Mathematical Methods for Digital Computers, Ralston A. and Wilf,H. S., (eds.), Wiley, New York.

[11] Glover, F. (1989). Tabu search-part I. ORSA Journal on computing, 1(3), 190-206.

[12] Glover, F. (1990). Tabu search—part II. ORSA Journal on computing, 2(1), 4-32.

[13] Glover, F., & Laguna, M. (2013). Tabu Search. (pp. 3261-3362). Springer New York.

[14] Harding, A., Shahbaz, M., & Kusiak, A. (2006). Data mining in manufacturing: A review. Journal of Manufacturing Science and Engineering, 128(4), 969-976.

[15] Hill, M. and Dixon, W.J., (1982) “Robustness in real life: A study of clinical laboratory data.”, Biometrics, 38:377-396.

[16] Hocking, R.R., (1976), “The Analysis and Selection of Variables in Linear Regression”, Biometrics, 32:1-49.

[17] Holland, J. H. (1975). Adaptation in natural and artificial system: an introduction with application to biology, control and artificial intelligence. Ann Arbor, University of Michigan Press.

[18] Huber, P. J. (1972). The 1972 wald lecture robust statistics: A review. The Annals of Mathematical Statistics, 1041-1067.

[19] Le Cam, L. (1986). The central limit theorem around 1935. Statistical science, 1(1), 78-91.

[20] Legendre, A. M., (1805), Nouvelles méthodes pour la détermination des orbites des comètes, Paris, p72-75.

[21] Marquardt, D.W., (1974), “Discussion (of Beaton and Tukey [1974])”, Technometrics, 15:189-192

[22] Micceri, T., (1989), “The unicorn, the normal curve, and other improbable creatures.”, Psychological Bulletin, 105:156-166.

[23] Murphy, A., (2000), Scientific Investment Analysis, Greenwood Press, pp100

[24] Myers, S., Majluf, N., (1984), “Corporate Financing and Investment Decisions When Firms Have Information That Investors Do Not”, Journal of Financial Economics, 13:187-221

[25] Nath, R., Rajagopalan, B., & Ryker, R., (1997),"Determining the Saliency of Input Variables in Neural Network Classifiers," Journal of Computers and Operations Research, 24,8:767-773

[26] Pfaffenberger, R., and Dinkel, J., (1978), "Absolute deviations curve-fitting: an alternative to least squares," in David, H (ed.), Contributions to Survey Sampling and Applied Statistics, New York: Academic Press, pp. 279-294

[27] Rousseeuw, P. J., & Leroy, A. M. (2005). Robust regression and outlier detection (Vol. 589). John Wiley & Sons.

[28] Statistical Abstract of the U.S., 1998.

[29] Starr, C., Rudman, R., Whipple, C., (1976), "Philosophical Basis for Risk Analysis," Annual Review of Energy, 1:629-662

[30] Stein, E. and Stein, J., (1991), "Stock Price Distributions with Stochastic Volatility: An Analytic Approach," The Review of Financial Studies, 4:727-752

[31] Wilcox, R. R., (1997), Introduction to Robust Estimation and Hypothesis Testing, Chestnut Hill, MA: Academic Press, pp 1.

[32] Wilson, H., (1978),"Least-squares versus minimum absolute deviations estimation in linear models," Decision Sciences, 9:322-335.

[33] Yourstone, S. A., & Zimmer, W. J. (1992). Non-normality and the design of control charts for averages. Decision Sciences, 23, 1099-1113.