An Investigation of the Median-Median Method of Linear Regression

Elizabeth J. Walters, Christopher H. Morrell, and Richard E. Auer
Loyola College of Maryland

Journal of Statistics Education Volume 14, Number 2 (2006), ww2.amstat.org/publications/jse/v14n2/morrell.html

Copyright © 2006 by Elizabeth J. Walters, Christopher H. Morrell, and Richard E. Auer, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Least squares line.

Abstract

Least squares regression is the most common method of fitting a straight line to a set of bivariate data. Another less known method that is available on Texas Instruments graphing calculators is median-median regression. This method is proposed as a simple method that may be used with middle and high school students to motivate the idea of fitting a straight line to data. The median-median line may also be viewed as a method that is not greatly affected by outliers (robust to outliers). Our paper briefly reviews the median-median regression method, considers various examples to compare the median-median line to the least squares line, and investigates the properties of the median-median line versus the least squares line using a simulation study.

1. Introduction

Two trends have had a great impact on statistics education. One trend involves the continual advancement in computer technology that has systematically increased the power and refinement of statistical data analysis. In the 1960’s, computers revolutionized statistical methodology. During the 1970’s, personal computers placed computing capability right on the desks of individual researchers. This evolution has continued and now grammar school students learn data skills literally in the palm of their hands using graphing calculators.

The second trend is the introduction of statistical ideas earlier in the school curriculum. Younger students are being exposed to the question of what truths lurk beneath the surface of data. These students are now being trained to use statistics as the primary tool for researching any topic within any field.

Evidence of both trends is seen when elementary and secondary school students use graphing calculators to perform data analysis on exams and homework assignments. Much of the early statistical training is being accomplished by teaching the basics of what has been termed exploratory data analysis (EDA).

As the first pioneer in exploratory data analysis, Tukey (1971, p. v) effectively defined EDA as “… looking at data to see what it seems to say. It concentrates on simple arithmetic and easy-to-draw pictures. It regards whatever appearances we have recognized as partial descriptions, and tries to look beneath them for new insights.” In his recommendations for what type of activity belongs in introductory statistics coursework, Hahn (1988, p. 271) said: “When we get to data analysis, we should stress graphical methods or exploratory data analysis, over formal statistical procedures.” Cobb and Moore (1997, p. 815) echoed the sentiments of Hahn saying: “Students like exploratory analysis and find that they can do it, a substantial bonus when teaching a subject feared by many. Engaging them early on in the interpretation of results, before the harder ideas come to their attention, can help establish good habits when you get to inference.”

Similar thinking was the basis for the Quantitative Literacy program of the mid-1980’s which was sponsored by the American Statistical Association (2005a) and the National Council of Teachers of Mathematics Joint Committee on the Curriculum in Statistics and Probability. Funded in part by the National Science Foundation, this program involves “… education workshops and written materials to help elementary and secondary school teachers make statistics more accessible to students … with an emphasis on graphical techniques” (American Statistical Association, 2005b). These written works were made available by Dale Seymour Publications through a series of four books. All four of these books focus on a “hands-on” approach that utilizes data very familiar to students from areas like sports, television, music, etc. One of these, Exploring Data by Landwehr and Watkins (1986), introduces students to a method of fitting a line through bivariate data using a simple application of the median.

One major goal of young statisticians as they learn exploratory data analysis is studying the relationship between two variables. When the data are bivariate with two numerical variables, this usually involves fitting a straight line through points on a scatterplot. The graphing calculator is programmed with two methods of line fitting for bivariate data. One method, fitting the least squares line, is a procedure well studied and routinely applied. But the median-median line, on the other hand, is not so well known or understood.

This paper considers the performance of these two line-fitting techniques as an aide to budding statisticians and their teachers as they encounter exploratory data analysis. We intend to illuminate the behavior of the relatively unknown median-median line with several data sets and through a small simulation study. Readers may compare how the median-median line reacts in various data settings to the performance of the least squares line.

2. Historical Evolution of the Median-Median Line

The median-median line traces back to a line-fitting approach proposed by Wald (1940). He suggested a very simple method where the points on a scatterplot are separated into a left half and a right half based on the median of the x-scores in a sample of bivariate data. The means of the x-scores and the y-scores are calculated using the data from only the left half of the scatterplot and then calculated using only the data on the right half. Concentrating on the two points, \( \bar{X} \) and \( \bar{Y} \), Wald proposed finding a line connecting these points that is then adjusted up or down to better fit the full array of points on the scatterplot. (The subscripts \( R \) and \( L \) denote which half of the data is used and no subscript implies the means are taken over the entire sample.) His line of fit has slope

\[
\beta = \frac{\bar{Y} - \bar{Y}}{\bar{X} - \bar{X}}
\]

and y-intercept \( \alpha = \bar{Y} - \beta \bar{X} \).

A similar procedure suggested by Nair and Shrivastava (1942) breaks up the points on a scatterplot into three regions with each region containing about the same number of points. The means of the x and y points in the left and right regions are used to find the slope of the line of fit much as Wald suggested.

Brown and Mood (1951) used the two-region approach but found the slope of the line of fit using medians in place of means. The primary advantage of using this measure of center comes from the median’s inherent ability to resist the strong effect of outliers. Most students of statistics know that the mean can be affected greatly by outliers since they are included with equal weight with the rest of the data in the sum of the scores. But the median takes on the same value whether the largest score in a data set is just somewhat larger than the rest of the data or is much larger than the second biggest score. In the context of fitting points on a scatterplot, this implies that a single point far from the general sloping trend of the rest of the points would not apply such a large “tug” on the location of the line of fit if the median is used to find the line.
Readers may recall that the least squares line is found by minimizing the sum of the squared distances that each point lies from the line. Since these distances are squared, Hartwig and Dearing (1979, p. 34) noted that “… cases lying farther and farther from the regression line increase the sum of the squared residuals at an increasing rate … [and the line] will have to come reasonably close to them to satisfy the least squares criterion and, therefore, the least squares regression line will lack resistance to the excessive influence of a few atypical cases.” This means that Brown and Mood’s method is not only simple to apply, but also has the advantage of not allowing outlying cases to have undue impact on determining the line of fit.

Like Brown and Mood, Tukey (1971) utilized the medians in finding his line of fit, but he did so borrowing the three-region approach of Nair and Shrivastava. His line of fit, called the resistant line, is considered a basic methodology of exploratory data analysis. The median-median line provides the first iteration in the procedure to find Tukey’s resistant line. To obtain Tukey’s resistant line, the residuals are used to adjust the parameters in an iterative fashion. Appendix A provides the algorithm used by Texas Instruments to compute the median-median line. Tukey’s resistant line may be obtained using Minitab under the EDA submenu of Stat.

For those who wish to learn more about the broad methodology of exploratory data analysis, Velleman and Hoaglin (1981) provide a gentle review of the entire subject. To learn more about the median-median line at a more sophisticated mathematical level, readers are encouraged to consider Emerson and Hoaglin (1983) and Johnstone and Velleman (1985).

3. Examples

In order to illustrate the comparative performance of the median-median line and the least squares line, we consider three example data sets from popular introductory statistics texts. The first contains no outliers. The second illustrates the impact of influential observations and outliers on the least squares line. The third example examines the data in two ways: one way has a single extreme influential observation while the other has a single outlier.

3.1 Manatees and Motor Boats

An investigation of the relationship between the number of manatees killed by boats in Florida and the number of powerboat registrations in that state, from 1977 to 1990 (Moore, 1997, p 347), shows a strong positive correlation with no obvious outliers or influential points ($r = 0.941$, $p$-value $< 0.001$). To calculate the equation of the median-median line, one first divides the data into three regions by $x$-score (in this case, the number of powerboat registrations, in thousands) (see Table 1). In the case when there are ties in the $x$-scores that would result in these points being in different regions, all points with the same $x$-score are placed in an outer region. This may result in the middle region having less than 1/3 of the observations (see Appendix A and Appendix B). While this does not occur with the data for Example 3.1, it does happen with the data for Example 3.2.

Table 1. Manatee data divided into three regions based on $x$-score. In each region, the first column $x$ is the number of powerboat registrations, in thousands, and the second column $y$ is the number of manatees killed.

| Region 1 | Region 2 | Region 3 |
|----------|----------|----------|
| $x$      | $y$      | $x$      | $y$      | $x$      | $y$      |
| 447      | 13       | 513      | 24       | 614      | 33       |
| 460      | 21       | 526      | 15       | 645      | 39       |
| 481      | 24       | 559      | 34       | 675      | 43       |
| 498      | 16       | 585      | 33       | 711      | 50       |
| 512      | 20       | 719      | 47       |

Once the data are divided into three regions, the median of the $x$- and median of the $y$-scores are calculated for each region. The resulting three points for these data are termed the median-median points $(\bar{x}_1, \bar{y}_1) = (481, 20)$, $(\bar{x}_2, \bar{y}_2) = (542.5, 28.5)$, $(\bar{x}_3, \bar{y}_3) = (675, 43)$. The slope of the median-median line is the slope of the line passing through $(\bar{x}_1, \bar{y}_1)$ and a parallel line passing through $(\bar{x}_3, \bar{y}_3)$. Moving the line connecting the two outer points one third of the way to the line through the point in the center region yields the median-median line (see Figure 1). Note that the line that connects the two outer points is based on two of the median-median points, while the line through $(\bar{x}_3, \bar{y}_3)$ is based on only this median-median point. Thus, choosing to move the first line one third of the way towards the second effectively makes the median-median line a weighted average of the outer two lines. The resulting $y$-intercept is $b = \frac{20 + 28.5 + 43}{3} = 29.833$. This yields the equation of the median-median line: $\hat{y} = -36.62285 + 0.11856x$. Note that the median-median and least-squares lines have similar equations, as is evident from the scatterplot (see Figure 2).
Figure 1. Illustration of the median-median method for the manatee data.
3.2 Mental Aptitude and Age at First Word

A scatterplot of scores on a test of mental aptitude (Gesell Adaptive Test) versus age at first word, for 21 children Moore and McCabe (2003, p. 161), shows a moderate negative correlation ($r = -0.640$, $p$-value = 0.002). Three points in this scatterplot are notable. Two children (●) began speaking at a much later age than the rest of the children. Another child (♦) scored much higher than we would expect, given his age at first word. The scatterplot in Figure 3 compares the median-median line to two least-squares lines: one for the entire set of data, and the other without the data for the latest-speaking child. Omitting the data point for this child leads to a least-squares line that is closer to the median-median line. Omitting the data for the unusually high-scoring child (●) increases the strength of the sample correlation but does not noticeably change the equation of the least-squares line; thus, that least squares line is not included in the scatterplot. The data are provided in Appendix B.
3.3 Meal Choices and Body Habits

A paper by Jiang and Hunt (1983) titled “The Relation Between Freely Chosen Meals and Body Habits” reports on the relationship between diet and body build. In particular, the researchers were interested in determining how the energy intake in one’s diet is related to that individual’s body build. Figure 4 plots dietary energy density (DED = kilocalories consumed divided by the weight of food eaten in grams) against the body mass index (BMI) (also known as the Quetelet index defined as weight in kg / (height in meters)^2) for nine individuals Devore and Peck (2001, p. 145-146). Including the outlying point (■) representing an individual with an unusually low energy density for his body build, these data have a moderate positive correlation ($r = 0.658$, $p$-value = 0.054). Note that the least-squares line calculated without the outlier differs greatly from the original least-squares line but is almost identical to the median-median line.
For illustration, we also consider DED as the explanatory variable and BMI as the response variable. In this scenario, the observation that was previously a high leverage point is now an observation with a large residual (Figure 5). Now the least squares line is not affected as much by the outlying observation as in Figure 4 where the observation has both high leverage and is far from the trend formed by the remaining observations. The data are provided in Appendix B.
4. Simulation Study

To compare the statistical properties of least squares and median-median estimates of the slope of a linear regression model, a simulation study is conducted that considers a variety of conditions. In every simulation, the underlying linear regression model is defined to be \( Y = 0 + 1X + \epsilon \). In order to create simulated \( Y \)-values, the expected \( Y \)-value is computed and then normal error is added. To create an outlier, 5 times the absolute value of the error is added to or subtracted from the expected \( Y \)-value at that particular point.

The conditions under which the simulation is performed are:

i. two sets of values of the explanatory variable:
   Set 1: \( X \)-values = \( (1, 2, 3, \ldots, 24) \), and
   Set 2: \( X \)-values = \( (2, 2, 4, \ldots, 24, 24) \),

ii. two levels of the error standard deviation (\( \sigma = 1 \) and 5),

iii. a number of outlier possibilities.

Without loss of generality, in most cases we assume that the outlier occurs in the upper region of \( X \)-values, that is, the outlier is one or two of \( Y_{17} \), \( Y_{23} \), \( Y_{24} \). Table 2 provides a summary of the outliers generated in the simulation study.

| Outlier Generation | Description |
|--------------------|-------------|
| a) \( Y_{13} \) = 0 + 1\( X_{13} \) + 5\( \epsilon_{24} \) | No Outliers |
| b) \( Y_{17} \) = 0 + 1\( X_{17} \) + 5\( \epsilon_{24} \) | One high outlier in the middle of the middle region |
| c) \( Y_{17} \) = 0 + 1\( X_{17} \) + 5\( \epsilon_{24} \) | Outliers at the first two \( X \)-values in the upper region, one high and one low |
| d) \( Y_{17} \) = 0 + 1\( X_{17} \) + 5\( \epsilon_{24} \) | One high outlier at the first \( X \)-value in the upper region |
| e) \( Y_{23} \) = 0 + 1\( X_{23} \) + 5\( \epsilon_{24} \) | Outliers at the two largest \( X \)-values, one high and one low |
| f) \( Y_{17} \) = 0 + 1\( X_{17} \) + 5\( \epsilon_{24} \) | High outliers at the first two \( X \)-values in the upper region |
| g) \( Y_{17} \) = 0 + 1\( X_{17} \) + 5\( \epsilon_{24} \) | A low outliers at the first \( X \)-value in the upper region and a high outlier at the largest \( X \)-value |
| h) \( Y_{24} \) = 0 + 1\( X_{24} \) + 5\( \epsilon_{24} \) | A single high outlier at the largest \( X \)-value |
| i) \( Y_{24} \) = 0 + 1\( X_{24} \) + 5\( \epsilon_{24} \) | High outliers at the two largest \( X \)-values |

For each type of outlier generation, 1000 samples are simulated. For each sample the least squares and median-median estimates of the intercept and slope are computed. Tables 3 to 6 provide the following statistics on the 1000 simulated least squares and median-median slope estimates: means, standard deviations, and mean square errors (variance of simulated slopes plus bias squared where the bias = mean - 1) for each of the nine methods of generating the data. If the mean of the simulated slope estimates is 1, the estimator is considered unbiased. The standard deviation provides a measure of spread of the estimates of the slopes. And the mean square error provides a measure of precision of how far the estimates vary from the true slope. Since the main focus of this paper is on the slopes of the estimated line, we do not consider the estimated intercepts in this discussion of the simulation results. These tables also show the leverage value (see Neter, Kutner, Nachtsheim, and Wasserman (1996, p. 375-377)) for the design points at which the outlier is generated. Leverage values help to identify outlying \( X \)-values that, in conjunction with extreme \( Y \)-observations, may lead to data points with a large influence on the slope and intercept of the line. It is recommended to compare the leverage value with \( \frac{2p}{n} \) where \( p \) is the number of parameters in the linear regression model. In our case \( p = 2 \) (for the intercept and slope) and \( n = 24 \) so \( \frac{2p}{n} = 0.167 \). In our example, no design point has a leverage that exceeds this value though the leverage of the most extreme \( X \)-values (1 and 24) come close to \( \frac{2p}{n} \).

Table 3. Monte Carlo Results: Means, standard deviations, and mean square errors (x 10^3) of the least squares (LS) and median-median (MM) slope estimates for 1000 replications.

Model: \( Y = 0 + 1X + \epsilon \); \( X \)'s: Set 1; \( \epsilon \sim N(0, 1^2) \).
Table 4. Monte Carlo Results: Means, standard deviations, and mean square errors ($10^3$) of the least squares (LS) and median-median (MM) slope estimates for 1000 replications.

Model: $Y = 0 + 1X + \varepsilon$; $X$'s: Set 1; $\varepsilon \sim N(0, 5^2)$.

| Type of Outlier | Mean Slope Estimate | Standard Deviation | MSE ($10^3$) |
|-----------------|---------------------|--------------------|--------------|
|                 | Leverage            | LS | MM | LS | MM | LS | MM |
| a) No outliers  |                     | 1.0017 | 1.0037 | 0.0295 | 0.0509 | 0.873 | 2.604 |
| b) One high in middle of middle region | 0.0419 | 1.0035 | 1.0037 | 0.0295 | 0.0509 | 0.882 | 2.604 |
| c) First two $X$-values in upper region, one high and one low | 0.0593 | 0.9980 | 1.0291 | 0.0349 | 0.0568 | 1.222 | 4.073 |
| d) One high at first $X$-value in upper region | 0.0593 | 1.0170 | 1.0295 | 0.0315 | 0.0564 | 1.285 | 4.051 |
| e) Two largest $X$-values, one high and one low | 0.1375 | 0.0680 | 1.0054 | 0.9703 | 0.0472 | 0.0557 | 2.257 | 3.985 |
| f) High at first two $X$-values in upper region | 0.0593 | 0.0608 | 1.0537 | 1.0571 | 0.0358 | 0.0569 | 2.417 | 6.521 |
| g) Low at first $X$-value in upper region, high at largest $X$-value | 0.0593 | 0.1567 | 1.0254 | 1.0037 | 0.0424 | 0.0511 | 2.442 | 2.625 |
| h) High at largest $X$-value | 0.1567 | 0.1567 | 1.0749 | 1.0047 | 0.0482 | 0.0513 | 7.933 | 2.654 |

Table 5. Monte Carlo Results: Means, standard deviations, and mean square errors ($10^3$) of the least squares (LS) and median-median (MM) slope estimates for 1000 replications.

Model: $Y = 0 + 1X + \varepsilon$; $X$'s: Set 2; $\varepsilon \sim N(0, 1^2)$.

| Type of Outlier | Mean Slope Estimate | Standard Deviation | MSE ($10^3$) |
|-----------------|---------------------|--------------------|--------------|
|                 | Leverage            | LS | MM | LS | MM | LS | MM |
| a) No outliers  |                     | 1.0087 | 1.0141 | 0.1473 | 0.1850 | 21.773 | 34.424 |
| b) One high in middle of middle region | 0.0419 | 1.0173 | 1.0141 | 0.1475 | 0.1850 | 22.056 | 34.424 |
| c) First two $X$-values in upper region, one high and one low | 0.0593 | 0.0680 | 0.9902 | 1.0614 | 0.1746 | 0.1948 | 30.581 | 41.717 |
| d) One high at first $X$-value in upper region | 0.0593 | 1.0848 | 1.0895 | 0.1576 | 0.1898 | 32.029 | 44.034 |
| e) Two largest $X$-values, one high and one low | 0.1375 | 0.0680 | 1.0272 | 0.9619 | 0.2362 | 0.1985 | 56.530 | 40.854 |
| f) High at first two $X$-values in upper region | 0.0593 | 0.0608 | 1.1786 | 1.1702 | 0.1691 | 0.1952 | 60.493 | 67.071 |
| g) Low at first $X$-value in upper region, high at largest $X$-value | 0.0593 | 0.1567 | 1.1271 | 1.0155 | 0.2118 | 0.1969 | 61.014 | 39.010 |
| h) High at largest $X$-value | 0.1567 | 0.1567 | 1.2023 | 1.0423 | 0.2028 | 0.1913 | 81.972 | 48.385 |
| i) High at two largest $X$-values | 0.1375 | 0.1567 | 1.3743 | 1.0888 | 0.2408 | 0.2009 | 198.085 | 48.246 |

Table 6. Monte Carlo Results: Means, standard deviations, and mean square errors ($10^3$) of the least squares (LS) and median-median (MM) slope estimates for 1000 replications.

Model: $Y = 0 + 1X + \varepsilon$; $X$'s: Set 3; $\varepsilon \sim N(0, 1^2)$.

| Type of Outlier | Mean Slope Estimate | Standard Deviation | MSE ($10^3$) |
|-----------------|---------------------|--------------------|--------------|
|                 | Leverage            | LS | MM | LS | MM | LS | MM |
| a) No outliers  |                     | 1.0018 | 1.0040 | 0.0296 | 0.0540 | 0.879 | 2.556 |
| b) One high in middle of middle region | 0.0425 | 1.0053 | 1.0040 | 0.0297 | 0.0504 | 0.910 | 2.556 |
| c) First two $X$-values in upper region, one high and one low | 0.0635 | 1.0015 | 1.0342 | 0.0350 | 0.0569 | 1.227 | 4.407 |
Table 6. Monte Carlo Results: Means, standard deviations, and mean square errors (× 10^3) of the least squares (LS) and median-median (MM) slope estimates for 1000 replications.

Model: Y = 0 + 1X + \( \varepsilon \); X's: Set 2; \( \varepsilon \sim N(0, 5^2) \).

**Lower region:** \( X_1, \ldots, X_8 \); **Middle region:** \( X_9, \ldots, X_{16} \); **Upper region:** \( X_{17}, \ldots, X_{24} \).

| Type of Outlier                                               | Leverage | Mean Slope Estimate | Standard Deviation | MSE (× 10^3) |
|---------------------------------------------------------------|----------|---------------------|--------------------|--------------|
| a) No outliers                                               |          | 1.0090              | 1.0138             | 0.1479       | 0.1849       | 21.955 | 34.378 |
| b) One high in middle of middle region                       | 0.0425   | 1.0263              | 1.0138             | 0.1485       | 0.1849       | 22.744 | 34.378 |
| c) First two X-values in upper region, one high and one low  | 0.0635   | 1.0076              | 1.0628             | 0.1751       | 0.1949       | 30.718 | 41.930 |
| d) One high at first X-value in upper region                 | 0.0635   | 1.0940              | 1.0876             | 0.1607       | 0.1898       | 34.660 | 44.698 |
| e) Two largest X-values, one high and one low                | 0.1474   | 1.0108              | 0.9632             | 0.2371       | 0.1988       | 56.333 | 40.874 |
| f) High at first two X-values in upper region                | 0.0635   | 1.1797              | 1.1700             | 0.1700       | 0.1942       | 61.192 | 66.614 |
| g) Low at first X-value in upper region, high at largest X-value | 0.0635 | 1.1112              | 1.0154             | 0.2114       | 0.1980       | 57.055 | 39.441 |
| h) High at largest X-value                                   | 0.1474   | 1.1952              | 1.0467             | 0.2001       | 0.1917       | 78.143 | 38.939 |
| i) High at two largest X-values                              | 0.1474   | 1.3763              | 1.0886             | 0.2427       | 0.2008       | 200.505 | 48.171 |

Figures 6 to 8 provide density estimates for three of these examples for the first set of x-values (1, 2, 3, ..., 24): no outliers, a moderate outlier example (high and low outliers at the end — Table 2, design (e)), and the most extreme outlier example (two high outliers at the end — Table 2, design (i)). The density estimates are simply the simulated distributions of the slope estimates and are computed from the 1000 slope estimates using the SPlus function density.

The simulation study comparing the performance of the two methods of regression when no outliers are present (Figure 6) shows that both the median-median and LS estimates are close to being unbiased. However, there is less variation in the least-squares slope estimates than in the median-median slope estimates.
Density Estimates of Distributions of Slopes
No Outliers

- Median-Median, $\sigma = 1$
- Least Squares, $\sigma = 1$
- Median-Median, $\sigma = 5$
- Least Squares, $\sigma = 5$
- True Slope, $\beta = 1$

Density

Slope Estimate

0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6
Figure 6. Density estimates of the slopes for the 1000 replications for least squares and median-median estimates and for \( \sigma = 1 \) and \( \sigma = 5 \); X’s: Set 1; no outliers.

In the moderate case (Figure 7, one high and one low outlier at the end – Table 2, design (e)), the least-squares method still tends to give unbiased estimates of the slope while the median-median estimates are slightly biased to the low side and exhibit smaller MSE than the LS line when \( \sigma = 5 \).
Density Estimates of Distributions of Slopes
1 high and 1 low outlier at the end - Table 2, design (e)
Figure 7. Density estimates of the slopes for the 1000 replications for least squares and median-median estimates and for $\sigma = 1$ and $\sigma = 5$; $X$’s: Set 1; one high and one low outlier at the end – Table 2, design (e).

In the extreme case (Figure 8, two high outliers at the end – Table 2, design (i)), however, the median-median method performs much better, with less variation and more accuracy than the least-squares method, especially in the case of greater population standard deviation ($\sigma = 5$ vs. $\sigma = 1$).
Density Estimates of Distributions of Slopes
Two high outliers at the end - Table 2, design (i)
5. Conclusion

The median-median regression line is intended as a simple, resistant alternative to the least-squares line. In particular, it provides a method of line fitting that is computationally accessible to young students of statistics, while at the same time giving estimates that are robust, or less sensitive, to outliers than those given by least-squares regression.

The three examples provided in this paper show that, when an influential point is present, the median-median line may be less affected by the influential point than the least-squares line. The simulation study further supports the argument for using the median-median line as a resistant method of regression when outliers are present, especially in the most extreme cases (one or two high outliers at the end; see Table 2, design (i)) and in cases with greater population variance when the outlier is on the most extreme x-value.

On the other hand, there exists a well-developed set of tools for inference and for detecting the presence of unusual observations when using least-squares regression under normal error assumptions. In contrast, inference about the median-median estimates would require bootstrapping to obtain approximate standard errors, confidence intervals/regions, and p-values for tests. Given this, in addition to the superior performance of least-squares estimation when either no outliers or moderate outliers are present, the median-median method of regression would not appear to be a valuable tool at either the professional or collegiate level.

However, in elementary and middle schools the median-median method of estimating a line may be a reasonable approach to describe a linear relationship in a scatterplot. The teacher may first introduce the idea of a scatterplot as a way of visualizing the association between a response and explanatory variable. If there is a straight-line trend in the plot, the straight line may be fit by eye using a ruler to try to capture the trend. While each student will likely be satisfied that their eye-balled line fits the data well, they will also recognize that each student would have chosen a slightly different line. This will suggest the need for a more objective and structured method of fitting a line. The median-median line will not only be simple to find, but may also feel very connected to the visual process each student had just undertaken.

Appendix A: Texas Instruments median-median algorithm

The following is the documentation provided by Texas Instruments on their web page on how the TI graphing calculators compute Median-Median estimates.

Median-Median Line Algorithm on Graphing Handhelds.

Solution:

What method is used to calculate the median-median line?

The goal of the median-median line is to:

1) Divide the data into three parts with an equal number of data points
2) Define a summary point for each part
3) Use the three summary points to define the median-median line

How the TI Calculator does this:

1) The calculator will attempt to break the list into three equal parts without breaking up data groups of equal x values. In this case, the algorithm we use is designed to include at least 1/3 of the points in the left and right groups. What ever is left over is put into the center group, hopefully the remaining 1/3. The approach chosen was to fill the outside groups first and allocate the remaining data points to the center group. If the center group is empty, an error message is generated.

The reason equally x values are not split up is to ensure the same results are produced independent of the order the data appears in the original list. Without this restriction, a different result could be produced depending on the ordering in the data input lists.

2) A summary point is simply the median of all x's and y's in that part. Let's call the summary points (x1,y1), (x2,y2), and (x3,y3).

3) The median-median line will be parallel to the line going through the points (x1,y1) and (x3,y3)

\[
a = \frac{(y3 - y1)}{(x3 - x1)}
\]

and 1/3 the distance between the line through the two summary points and a parallel line going through the second summary point (x2,y2).

\[
b = \frac{(y1 + y2 + y3 - a(x1 + x2 + x3))}{3}.
\]

An advantage of the median-median line over a least-squares line, is that stray data points do not affect the end result very much.

Appendix B: Data sets used in the examples

Example 1. Manatee Deaths and Power Boat Registrations illustrating the 3 groups (data used with permission of W. H. Freeman and Company)

| Power Boat Registrations | Manatee Deaths |
|--------------------------|----------------|
| 447                      | 13             |
| 460                      | 21             |
| 481                      | 24             |
| 498                      | 16             |
| 512                      | 20             |
Example 2. Gessell Scores and Age illustrating the 3 regions (data used with permission of W.H. Freeman and Company)

| Age (x) | Gessell Score (y) |
|---------|-------------------|
| 7       | 113               |
| 8       | 104               |
| 9       | 91                |
| 9       | 96                |
| 10      | 83                |
| 10      | 83                |
| 10      | 100               |
| 10      | 100               |
| 11      | 100               |
| 11      | 84                |
| 11      | 102               |
| 11      | 86                |
| 12      | 105               |
| 15      | 95                |
| 15      | 102               |
| 17      | 121               |
| 18      | 93                |
| 20      | 87                |
| 20      | 94                |
| 26      | 71                |
| 42      | 57                |

Example 3. Body Mass Index (BMI) and Dietary Energy Density (DED) illustrating the 3 regions (data used with permission of Thompson Learning)

| BMI (x) | DED (y) | DED (x) | BMI (y) |
|---------|---------|---------|---------|
| 21.1    | 0.54    | 0.44    | 21.5    |
| 21.5    | 0.44    | 0.54    | 21.1    |
| 22.1    | 0.67    | 0.67    | 22.1    |
| 22.3    | 0.78    | 0.78    | 22.3    |
| 22.4    | 0.90    | 0.86    | 22.8    |
| 22.8    | 0.86    | 0.90    | 22.4    |
| 23.1    | 0.91    | 0.91    | 23.1    |
| 23.3    | 0.94    | 0.93    | 26.8    |
| 26.8    | 0.93    | 0.94    | 23.3    |

References

American Statistical Association (2005a), *Quantitative Literacy – An Overview* from www.amstat.org/education/index.cfm?fuseaction=QL.workshops

American Statistical Association (2005b), *Materials for K-12 Statistics Education* from www.amstat.org/education/index.cfm?fuseaction=k12material

Brown, G. W. and Mood, A. M. (1951), “On Median Tests for Linear Hypotheses,” *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA: University of California Press, 159-166.

Cobb, G. W. and Moore, D. S. (1997), “Mathematics, Statistics, and Teaching.” *The American Mathematical Monthly*, 104(9), 801-823.
Acknowledgements

We thank the editor and referees who provided comments that greatly improved the manuscript.

Elizabeth J. Walters
Mathematical Sciences Department
Loyola College in Maryland
Baltimore, MD 21210-2699
U.S.A.
ewalter@loyola.edu

Christopher H. Morrell
Mathematical Sciences Department
Loyola College in Maryland
Baltimore, MD 21210-2699
U.S.A.
chm@loyola.edu

Richard E. Auer
Mathematical Sciences Department
Loyola College in Maryland
Baltimore, MD 21210-2699
U.S.A.
reu@loyola.edu