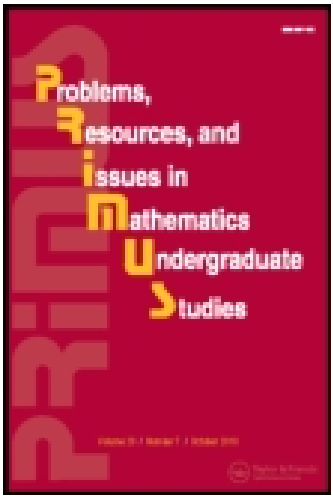

PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/upri20

Using Quartile-Quartile Lines as Linear Models

Sheldon P. Gordon

Accepted author version posted online: 03 Jan 2015.

To cite this article: Sheldon P. Gordon (2015) Using Quartile-Quartile Lines as Linear Models, PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies, 25:5, 389-399, DOI: 10.1080/10511970.2014.993053

To link to this article: http://dx.doi.org/10.1080/10511970.2014.993053

PRIMUS, 25(5): 389–399, 2015

ISSN: 1051-1970 print / 1935-4053 online
DOI: 10.1080/10511970.2014.993053

Taylor & Francis
Taylor & Francis Group

Using Quartile–Quartile Lines as Linear Models

Sheldon P. Gordon

Abstract: This article introduces the notion of the quartile–quartile line as an alternative to the regression line and the median–median line to produce a linear model based on a set of data. It is based on using the first and third quartiles of a set of (x, y) data. Dynamic spreadsheets are used as exploratory tools to compare the different approaches and to investigate the effects of sample size on the lines that are produced to fit random samples.

Keywords: Quartile–quartile line, regression line, median–median line, R^2 value, random samples.

1. INTRODUCTION

Both in mathematics and across most other quantitative disciplines, the most common interpretation of the line that is the best fit to a set of (x, y) data, and hence the best linear model, is the regression, or least-squares, line. It is based on the idea of finding the one line that comes closest to all of the data points in the sense that the sum of the squares of the vertical distances between the points and the line is a minimum.

However, one of the major drawbacks of the regression line is the fact that the sum of the squares is very sensitive to the presence of outliers: data points that are relatively far from the line. Any such points significantly contribute to the sum of the squares and hence have a disproportionate effect on the parameters of the regression line.

A pedagogical drawback is the fact that the usual derivation of the equation of the regression line involves an application of multivariate calculus—optimizing a function of two variables—although [3] presents a derivation based on algebra at the precalculus level.

Address correspondence to Sheldon P. Gordon, Department of Mathematics, Farmingdale State College of New York, Farmingdale, NY 11735, USA. E-mail: gordonsp@farmingdale.edu

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/upri

One alternative to the regression line that is used in many algebra and precalculus classes, particularly in high school classes, is the median–median line, which is programmed into the curve-fitting routines of most graphing calculators. Like the median itself, the median–median line is usually not affected by a few outliers and so is much more robust than the regression line. Also, since use of the median is stressed in modern statistics education, it seems a natural way to approach the idea of creating a line to fit a set of data. Unfortunately, although the rationale behind the median–median line can be fairly easily explained to students, it is not something that is easy to calculate by hand, either at the secondary or the beginning college level. Instead, it is likely that it is often applied either without showing the students how to do it, let alone having them do it themselves at least once.

Before continuing, we briefly review the rationale used for the TI calculator routine for the median–median line.

1. First, the data set $(x_k, y_k), k = 1, \ldots, n$, is ordered so that the x-values are in increasing order; each of the corresponding y-values is just "carried along."
2. Then the data set is subdivided into three groups having roughly the same number of points. The first group contains roughly the one-third of the data points with the smallest x-values along with the associated y-values. The third group contains roughly the one-third of the data points with the largest x-values, and the middle group contains the remaining data points. The first and third groups are formed in such a way that they have the same number of entries (equal to INT$((n + 2)/3 + 0.5)$, where INT is the greatest integer function). For example, if there are $n = 20$ data points, seven points would be assigned to the first and third groups and the remaining six points would be assigned to the middle group.
3. The medians of the x- and y-values in each of the three groups are calculated separately; denote them by (X_1, Y_1), (X_2, Y_2), and (X_3, Y_3), respectively.
4. The line through (X_1, Y_1) and (X_3, Y_3) is calculated.
5. This line is then translated one-third of the perpendicular distance D from the line toward the median point (X_2, Y_2) of the middle group while keeping the slope the same (see Figure 1). The resulting line is the median–median line.

Personally, the present author would not expect most of his students in an introductory college statistics course to understand this process to the extent of being able to use the ideas effectively with relatively large data sets without the use of the calculator routine. Although many students would follow the geometric reasoning, the number of calculations entailed in implementing the procedure by hand or with simple computational support likely would be unreasonable. This is compounded by the fact that applying the process requires a fairly large set of data (because of the need to partition the data into three

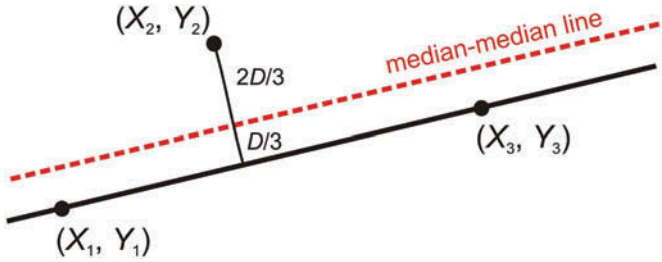

Figure 1. The median-median line.

groups and find the median of the x-values in each group), unlike the least-squares line that can be found based on only a handful of points. Thus, the use of the median–median line likely turns into a button-pushing routine with little underlying understanding beyond the fact that one is trying to find a line that closely fits to a set of data.

In this article, we present a third approach for constructing a line to fit a set of data, one based on the quartiles in the data, that is simpler to explain and considerably easier to implement by hand than the median–median line. It also appears to be comparably as effective in fitting a line to data as is either the least-squares line or the median–median line. Finally, it also has the advantage of building on ideas and methods (such as finding the equation of a line, interpreting the slope, and the use of the various summary statistics of a set of data) that students will have seen previously.

2. THE QUARTILE–QUARTILE LINE

Rather than focusing on the median, we look at the first and third quartiles of the data. In particular, we find these two quartiles for the x-values and the y-values separately. They can be found under the STAT menu of any TI graphing calculator (select CALC and then the one-variable statistics option for each list). Alternatively, with Excel, assuming there are, say, 50 data points in columns A and B, the quartiles can be determined by the formulas "=quartile(A1:A50,1)" for the first quartile and "=quartile(A1:A50,3)" for the third quartile of the x-values and comparable formulas for the quartiles of the y-values. Suppose we call the resulting points (X_1, Y_1) and (X_3, Y_3). We call the line that passes through these two points the *quartile–quartile line* or perhaps the *InterQuartile Line*. Finding its equation is a simple application of the point-slope formula for a line, a very important linkage between algebra and statistics, which helps integrate the statistical ideas into an algebra course (and vice versa) in a very natural way. (Note that the author could not find any reference to a quartile–quartile line, so the ideas here may be new.)

Some readers might wonder why we find the first and third quartiles of *both* the xs and the ys. First, quartiles often are not exactly equal to one of the data values, so there would be no corresponding value for the other variable. Second, even if the x-quartile exactly matches one of the x-values, there might be several corresponding y-values, and so there will not be a unique value to be used. Third, even if there is a unique y-value, that value might be an outlier and so would distort the resulting line. (Realize that the same difficulties can arise with the median–median line; it is just that these eventualities presumably are taken into account in the process programmed into the calculator routine.)

There is one complication with the quartile–quartile line. If the data values are trending upward, then the slope of the quartile–quartile line should be positive. If the data values are trending downward, the slope should be negative. However, the usual expression for the slope

$$m = \frac{Y_3 - Y_1}{X_3 - X_1}$$

will always be positive because the third quartile Y_3 is always greater than Y_1. Therefore, it is always necessary to look at the scatterplot of the data to decide on whether the trend is increasing or decreasing and then decide on whether the value for the slope should be positive or negative. However, this is not necessarily a bad step for students because it repeatedly reinforces several fundamental ideas. Perhaps then, it might be better to write the slope of the quartile–quartile line as

$$m = \begin{cases} \dfrac{Y_3 - Y_1}{X_3 - X_1} & \text{if the trend is upward} \\ -\dfrac{Y_3 - Y_1}{X_3 - X_1} & \text{if the trend is downward} \end{cases}$$

to emphasize the need to make a judgment call.

There are two natural questions. First, How well does this quartile–quartile line fit the data? Second, How does it compare to either the least-squares line or the median–median line? To investigate these questions, the author has developed a dynamic interactive spreadsheet in Excel [1] that generates a single random sample from an underlying bivariate population with a choice of sample size between $n = 4$ and $n = 40$, calculates the least-squares (regression) line, the median–median line, and the quartile–quartile line for the sample, and displays the results both graphically and numerically. It is available to readers and their students to dynamically investigate the ideas discussed here either in class or independently. The results of two different samples of size $n = 4$ are shown in Figure 2(a) and 2(b); comparable results for a sample of size $n = 24$ are shown in Figure 3. The data points are shown with the smaller (black)

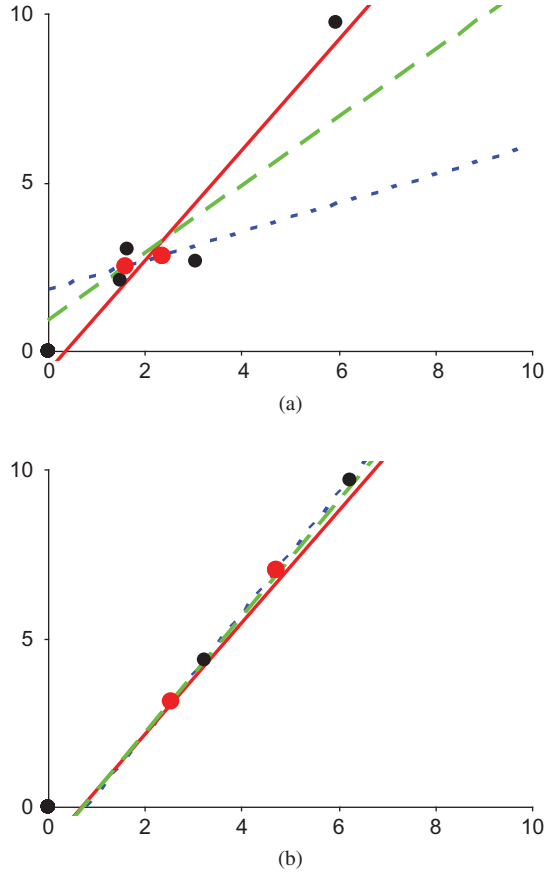

Figure 2. Two different results with $n = 4$ random points.

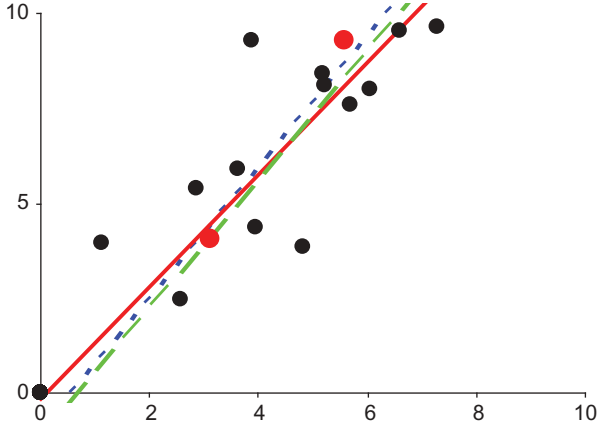

Figure 3. One sample with $n = 24$.

dots, whereas the median points used for the median–median line are the larger (red) dots.

Notice that the three lines in Figure 2(a) are quite different, but those in Figure 2(b) are extremely close. The corresponding equations of the three lines in Figure 2(a) and the associated R^2 values are

$$y = 1.63x - 0.56 \text{ (least squares)}, \quad R^2 = 0.926 \text{ (solid line)}$$

$$y = 0.42x + 1.85 \text{ (median} - \text{median)}, \quad R^2 = 0.845 \text{ (dotted line)}$$

$$y = 1.00x + 0.93 \text{ (quartile} - \text{quartile)}, \quad R^2 = 0.916 \text{ (dashed line)}$$

R^2, the coefficient of determination, is a generalization of the r^2 value that indicates how much of the variation in the ys is attributable to the linear model. The closer R^2 is to unity, the more the model accounts for the variations in y. For instance, when $R^2 = 0.926$, then 92.6% of the variation is explained by the linear model.

For the three lines in Figure 2(b), the equations and R^2 values are, respectively

$$y = 1.65x - 1.10, \quad R^2 = 0.996 \text{ (solid line)}$$

$$y = 1.79x - 1.43, \quad R^2 = 0.983 \text{ (dotted line)}$$

$$y = 1.72x - 1.24, \quad R^2 = 0.992 \text{ (dashed line)}$$

Notice that, in each case, the value of R^2 for the least-squares line is the greatest, followed by the quartile–quartile line, and then the median–median line. Based on the author's experience of looking at a great many sets of random samples using the Excel simulation, this appears to be fairly typical of what happens, however, exceptions certainly occur. With the simple click of a button, the spreadsheet displays the results for a new sample, so it is possible to generate many different samples. In each case, the R^2 value for the least-squares line is the largest; in the author's experience with many runs of the simulation, considerably more often than not, the value for the quartile–quartile line is greater than that for the median–median line, though usually the latter two are reasonably close. In most instances, the three values are reasonably close to one another.

The formula for R^2 is

$$R^2 = 1 - \frac{\sum_{k=1}^{n} (y_k - ax_k - b)^2}{\sum_{k=1}^{n} (y_k - \bar{y})^2},$$

where a and b are the parameters of the line $y = ax + b$ in question, the numerator is the sum of the squares of the vertical distances from the points to that line, and the denominator, which is the sum of the squares of the vertical distances from the y-values to their average $\bar{y}$, is the same for each line. Since the sum of the squares is minimal for the least-squares line, we should expect its R^2 value to be largest. Also, the further that points lie vertically away from any line, the greater the contribution of the squared difference in the numerator, and hence the greater the value of R^2.

Furthermore, as the sample size increases, the three lines appear to be more consistent with one another. For instance, with $n = 24$, notice that the three lines in Figure 3 are extremely close to one another, even though there is still a considerable amount of spread in the data points. The corresponding equations and R^2 values are, respectively

$$y = 1.48x - 0.14, \quad R^2 = 0.817 \text{ (solid line)},$$

$$y = 1.73x - 0.99, \quad R^2 = 0.762 \text{ (dotted line)},$$

$$y = 1.69x - 1.18, \quad R^2 = 0.769 \text{ (dashed line)}.$$

Repeated samples of the same size indicate that the above conclusions seem to be typical. For small sample sizes, there can be rather extensive differences among the three lines, but as the sample size increases, the three lines tend to be closer together. Perhaps a more telling display is based on repeated random samples where the quartile–quartile line of each sample is drawn. In Figure 4, we show 25 different quartile–quartile lines based on random samples of size $n = 4$. Notice the extent to which most of the lines differ from one another; this

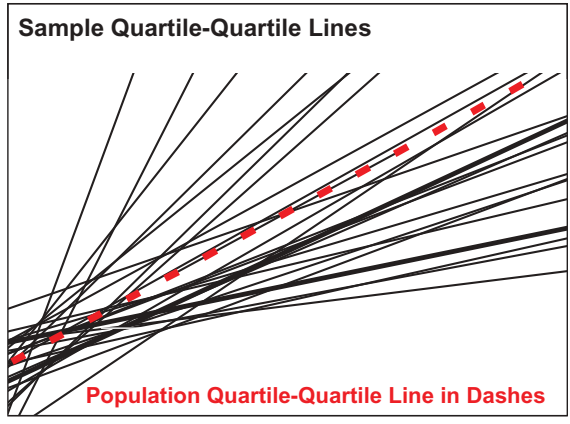

Figure 4. Samples with $n = 4$.

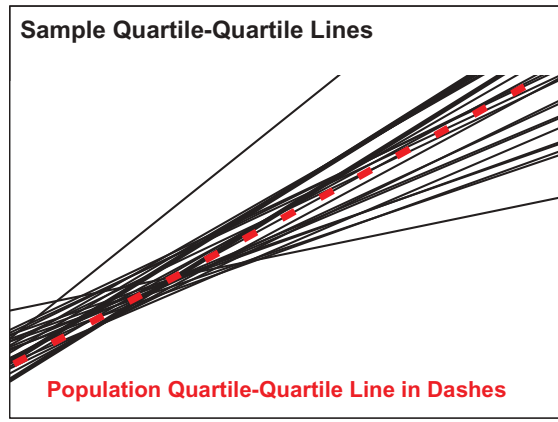

Figure 5. Samples with $n = 15$.

is due to the choice of the four points, although many tend to be reasonably similar to the population quartile–quartile line shown dotted in red. In Figure 5, we similarly show 25 sample quartile–quartile lines based on random samples of size $n = 15$ points. Observe how much closer all of the sample lines are to one another, as well as how close most of them are to the population quartile–quartile line. In this case, the effect of any single outlier is relatively minimal. We note that this Excel simulation [2] is also available to readers who want to investigate these ideas in more depth by looking at many different examples to gather information from which to judge the suitability of each approach.

We note that the results here with the sample quartile–quartile lines essentially parallel what happens with sample least-squares lines, though, if anything, the lack of consistency among the latter tends to be considerably more extreme for small sample sizes. Thus, the quartile–quartile line seems to be somewhat more consistent from one sample to another.

Figure 6 shows the scatterplot of the underlying population used in both Excel simulations; notice that the pattern is fairly linear. When the sample size is small, it is clear that some of the samples could contain points that do not fall into a pattern that mirrors the underlying bivariate population, and the resulting lines will have slopes that are distinctly different from the regression line for the population. However, as the sample size increases, it is less likely that this will occur and most, if not all, of the sample lines will be close to one another. This should be the case whenever the underlying population is roughly linear.

However, if the underlying population is very nonlinear—say U-shaped—then things could significantly change. The median–median line would likely be a better fit, since it takes the central third of the sample data into account whereas the quartile–quartile line does not do so. However, in such a case, one should not be fitting a line to the data anyway.

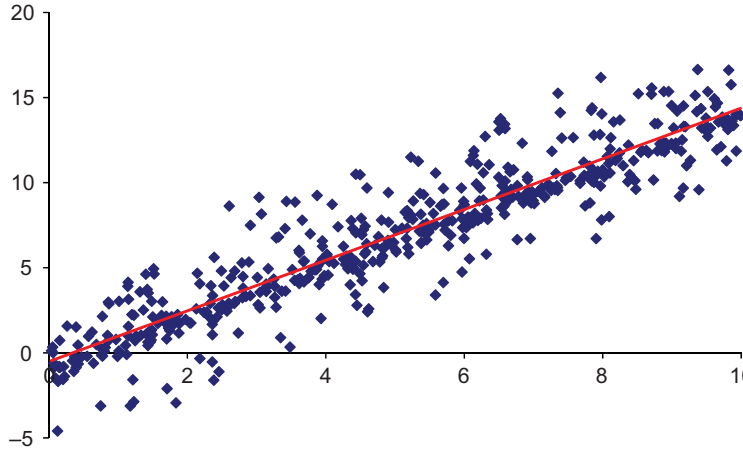

Figure 6. The underlying bivariate population and its regression line.

2.1. Some Limitations

There are several potential limitations to using the quartile–quartile line, including the uncertainty about the sign of the slope when it is close to zero or when the underlying population is decidedly nonlinear. Another limitation involves the choice of points for determining the slope. The quartile–quartile line uses points that are essentially one-quarter of the way in from either extreme and so may not closely reflect what happens at either end of the data. On the other hand, the quartile–quartile line is not influenced by outliers that can occur at either end of the data, unlike what can happen with the least-squares line. Note that the median–median line uses the medians of the upper and lower thirds of the data, and so is only about one-sixth of the way from either extreme

For that matter, some people use least squares in the sense of the perpendicular distance from each of the data points to a line or even the horizontal distance from the points to the line instead of the more usual vertical distances. Given the wide diversity of ways in which people define the best line to fit data, it is clear that there is no single universally accepted approach. As such, it is reasonable to suggest a variety of possibilities in class and to discuss with students some of the potential benefits and limitations of each approach, as well as the significance of the slope and intercept of each, along with the R^2 value. Such a discussion, pointing out that the decision of what the best fit means is a judgment issue that may depend on the particular data set, can be a very valuable learning experience. Students need to be reminded that using mathematics effectively involves understanding and thinking, not just the rote application of a collection of rules or the mindless pushing of a button.

2.2. Statistical Issues

There are two major issues that underlie all statistical processes. One is the effect of sample size on the results; the other is variability, either *within* a given sample or *between* different samples. For instance, picture students conducting a lab experiment in one of their science courses; the particular sample each student obtains is just one of a huge number of possible samples based on other sets of measurements. How does that one sample compare to the other possibilities? It is essential that these issues be brought front and center for students at all levels or we face the prospect that statistical education reduces to mindless button-pushing with little or no statistical understanding. The use of statistical simulations is a great way to investigate the effects of sample size and the variability between samples, as we have demonstrated above. Repeated random samples quickly demonstrate what can happen as students see the results of many different samples. They see clearly the effects of increasing the sample size in terms of increasing consistency and vice versa. Unfortunately, the random nature of the samples generated makes it much harder to examine the changes that occur *within* a sample, say by eliminating outliers. However, that lesson can be addressed in other ways that we will not go into here.

3. CONCLUSIONS

The quartile–quartile line seems to be comparable to both the least-squares line and the median–median line in its ability to serve as a linear model to fit a set of data that fall into a roughly linear pattern. However, the fact that it is computationally much simpler than the median–median line and perhaps slightly conceptually simpler makes it an attractive alternative to the latter, particularly in introductory courses.

REFERENCES

1. Gordon, S. P. Regression simulation, (currently item #3 on website), http://www.farmingdale.edu/faculty/sheldon-gordon/dynamicprecalculus.htm. Last accessed 4 February 2015.
2. Gordon, S. P., Comparing Lines that fit data, (currently item #6 on website), http://www.farmingdale.edu/faculty/sheldon-gordon/dynamicprecalculus.htm. Last accessed 4 February 2015.
3. Gordon, S. P. and F. S. Gordon. 2004. Deriving the regression equations without calculus. *Mathematics and Computer Education*. 38: 64–68.

BIOGRAPHICAL SKETCH

Sheldon P. Gordon is SUNY Distinguished Teaching Professor of Mathematics at Farmingdale State College. He has served on a number of national committees involved in undergraduate mathematics education with special emphasis on efforts to rethink calculus and courses below calculus. He is a co-author of *Functions, Data and Models: An Applied Approach to College Algebra, Functioning in the Real World* and *Contemporary Statistics: A Computer Approach* and a co-editor of the MAA Notes volumes, *Statistics for the Twenty First Century* and *A Fresh Start for Collegiate Mathematics: Rethinking the Courses Below Calculus.* He was also an original co-author of the texts developed under the Calculus Consortium based at Harvard.