THE SHORTEST DISTANCE IN DATA ANALYSIS

ALAN M. SAFER, KAGBA SUARAY, AND SALEEM WATSON

Abstract. By representing a set of N data points as a vector $\mathbf{x}$ in $\mathbb{R}^N$, we show that certain data analysis concepts, in particular regression and quantile regression, can be interpreted as vectors that minimize the distance to the vector $\mathbf{x}$, with respect to an appropriate metric or quasimetric.

1. The Metric is the Message

Many of the concepts in data analysis involve ideas of closeness or distance. For example, statistics textbooks introduce standard deviation as a measure of how far the data are from the mean, or the median as the middle of the data, and for two-variable data the regression line is described as the line closest to all the data points [5]. The words how far, middle, and closest suggest that some method of measuring distance, or some type of metric is being used. Attempts to axiomatize our intuitive ideas about distance include the notions of metrics, psudometrics, semimetrics, quasimetrics, proximity, uniformity, and others. Each of these definitions is intended to mathematically prescribe some intuitive notion of distance. In [4] it is shown that for one-variable data the mean, median, mode, and midrange are realized as values that minimize a distance in an appropriate metric. In this article we examine how metrics and quasimetrics can be used to put different regression methods in a metric context.

Recall that a *metric* on a set X is a function $d\colon X\times X \to [0,\infty)$ satisfying the following conditions. For every x, y, $z \in X$,

(i) $d(x,y) = 0$ if and only if $x = y$.
(ii) $d(x,y) = d(y,x)$.
(iii) $d(x,z) \le d(x,y) + d(y,z)$.

In this article we consider metrics on $\mathbb{R}^N$. So let $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ be vectors in $\mathbb{R}^N$. The standard (or Euclidean, or ℓ^2) metric d on $\mathbb{R}^N$ is defined by

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{N}(x_i - y_i)^2\right)^{1/2}. \tag{1}$$

We also consider the ℓ^1 metric (sometimes called the taxicab metric) defined by

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} |x_i - y_i| . \tag{2}$$

A *quasimetric* is defined just like a metric except that property (ii) of a metric is not required [6]. So the definition of a quasimetric leaves open the possibility that $d(x, y)$ is not equal to $d(y, x)$. A well-known example of a quasimetric space is a hilly country with several villages, where the distance between villages is taken to be time of travel. So the distance between a village at the top of a hill and one in the adjacent valley is not symmetric—it takes longer to travel uphill than downhill. In this article we use a particular quasimetric (called the tilted absolute value function) which weights distances between data points and a proposed regression line differently, depending on whether the points are above or below the line. We'll see that finding the line of best fit with respect to such a quasimetric gives the researcher the choice of lifting (or lowering) the line of best fit to be closer to those data points (points in a particular quantile) that a researcher may consider more significant for a particular study.

For each p, $0 < p < 1$, define the p-tilted absolute value function λ_p as follows [2]:

$$\lambda_p(t) = \begin{cases} (p-1)t, & \text{if} \quad t < 0; \\ pt, & \text{if} \quad t \geq 0. \end{cases} \tag{3}$$

The graph in Figure 1 explains the term p-tilted.

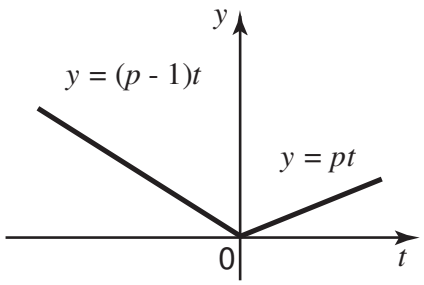

Figure 1. The tilted absolute value.

Lemma 1.1. *For each p, $0 < p < 1$, the function $q\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined by $q(x, y) = \lambda_p(x - y)$ is a quasimetric on $\mathbb{R}$.*

Proof. Property (i) of a metric is clearly satisfied. To show that property (iii) is satisfied let $x, z \in \mathbb{R}$ with $x < z$. If y is any other real number there are three cases to consider: $y < x < z$, $x < y < z$, and $x < z < y$. In the first case $y < x < z$, so

$$\begin{aligned} q(x,y) + q(y,z) &= p(x-y) + (p-1)(y-z) = p(x-z) - (y-z) \\ &\geq p(x-z) - (x-z) = (p-1)(x-z) = q(x,z). \end{aligned}$$

In the second case $x < y < z$, so

$$\begin{aligned} q(x,y) + q(y,z) &= (p-1)(x-y) + (p-1)(y-z) \\ &= (p-1)(x-z) = q(x,z). \end{aligned}$$

In the third case $x < z < y$, so

$$\begin{aligned} q(x,y) + q(y,z) &= (p-1)(x-y) + p(y-z) = p(x-z) - (x-y) \\ &\geq p(x-z) - (x-z) = (p-1)(x-z) = q(x,z). \end{aligned}$$

So inequality (iii) holds in all cases. □

For the vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ define

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} \lambda_p(x_i - y_i). \tag{4}$$

This definition is analogous to (2) except that the absolute value has been replaced by the p-tilted absolute value. In (4) each term has been weighted, the weight depending on whether the data point x_i is greater than or less than y_i. Although (2) and (4) are analogous, the situations are quite different in that (2) defines a metric but (4) does not. Indeed, (4) defines a quasimetric on $\mathbb{R}^N$ since it is a sum of quasimetrics.

2. Central Tendency and Quantiles

Consider a set of one-variable data $\{x_1, x_2, \ldots, x_N\}$ and suppose that we are interested in finding a measure of central tendency. Let us represent the data by a vector $\mathbf{x}$ in $\mathbb{R}^N$ and let $\mathbf{m}$ be a constant vector:

$$\mathbf{x} = (x_1, x_2, \ldots, x_N) \qquad \text{and} \qquad \mathbf{m} = (m, m, \ldots, m).$$

Intuitively, we would like the central tendency vector $\mathbf{m}$ to be the vector "closest" to $\mathbf{x}$; that is, we would like to find the value of $\mathbf{m}$ that minimizes the function

$$f(m) = d(\mathbf{x}, \mathbf{m}). \tag{5}$$

It is shown in [4] that if d is the ℓ^1 metric then the value of $\mathbf{m}$ that minimizes (5) is a median of the data. This value is not unique; when N is even, any value of m between the middle two data points minimizes (5).

The median is a special case of a *quantile*. Other commonly used quantiles are the quartile, quintile, decile, or percentile. In general, a pth quantile is a value such that the proportion of the data below that value is at most p and the proportion of the data above that value is at most $1-p$ [1]. We show that a pth quantile, just like a median, is determined as a solution to a shortest distance problem.

Theorem 2.1. *Let* $\mathbf{x}$ *be a data vector and let* $\mathbf{m}$ *be a constant vector in* $\mathbb{R}^N$. *The value of* m *that minimizes the quasimetric distance*

$$d(\mathbf{x},\mathbf{m}) = \sum_{i=1}^{N} \lambda_p(x_i - m)$$

is a p*th quantile.*

Proof. Let

$$f(m) = \sum_{i=1}^{N} \lambda_p(x_i - m) = \sum_{x_i<m} (p-1)(x_i - m) + \sum_{x_i>m} p(x_i - m).$$

Taking the derivative with respect to m and setting it equal to zero we get

$$f'(m) = \sum_{x_i<m} (1-p) + \sum_{x_i>m} (-p) = 0. \tag{6}$$

Let r be the number of data points for which $x_i < m$ and s the number of data points for which $x_i > m$. So $r+s \le N$. Then (6) becomes

$$f'(m) = r(1-p) + s(-p) = 0.$$

Solving we get $p = r/(r+s)$. Now the proportion of the data that is less than m is $r/N \le r/(r+s) = p$, and the proportion of the data that is greater than m is $s/N \le s/(r+s) = 1 - r/(r+s) = 1-p$. Thus the function f is minimized when m is a pth quantile. □

3. Linear Regression

A set of two-variable data is a set of the form $\{(x_1,y_1),(x_2,y_2),\ \ldots,\ (x_N,y_N)\}$. The goal in linear regression is to find a linear function $y = ax+b$ whose graph is as close as possible to all the data points. Write $\widehat{y}_i = ax_i + b$ and consider the vectors $\mathbf{y}$ and $\mathbf{L}$ in $\mathbb{R}^N$ given by

$$\mathbf{y} = (y_1, y_2, \ldots, y_N) \qquad \text{and} \qquad \mathbf{L} = (\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_N). \tag{7}$$

It is well-known that the values of a and b for the regression line $y = ax+b$ are those that minimize the distance $d(\mathbf{y},\mathbf{L})$ where d is the ℓ^2 metric on

$\mathbb{R}^N$ described in (1). This distance is the function F of the variables a and b given by

$$F(a,b) = \sum_{i=1}^{N} |y_i - (ax_i + b)|^2 = \sum_{i=1}^{N} |y_i - \widehat{y}_i|^2 .$$

So the regression line is the line that minimizes the sum of the squares of the *residuals* $y_i - \widehat{y}_i$. The graph in Figure 2 contains some data points and the corresponding regression line, with the magnitudes of the residuals represented by dashed line segments. Formulas for the values of a and b that minimize the function F are easily derived by using calculus or linear algebra.

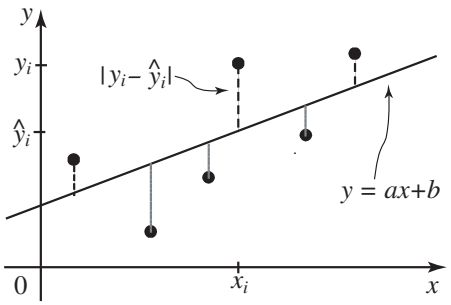

Figure 2. Regression Line and Residuals.

We now consider the ℓ^1 metric as a different method of measuring distance for the purpose of finding a line of best fit for two-variable data. The ℓ^1 regression line, or *median regression line* [1], is obtained by finding the linear function $f(x) = ax + b$ that minimizes the distance $d(\mathbf{y}, \mathbf{L})$ between the vectors $\mathbf{y}$ and $\mathbf{L}$ defined in (7) in the ℓ^1metric on $\mathbb{R}^N$. This distance is the function F of the two variables a and b given by

$$F(a,b) = \sum_{i=1}^{N} |y_i - (ax_i + b)| = \sum_{i=1}^{N} |y_i - \widehat{y}_i| .$$

So the median regression line is the line that minimizes the sum of the absolute values of the residuals $y_i - \widehat{y}_i$ (see Figure 2). In this case, finding formulas for the values of a and b that minimize the function F presents formidable difficulties. The presence of the absolute value makes the calculus approach intractable. Since the ℓ^1 metric is not derived from an inner product, the linear algebra approach is also not workable.

It is apparent however that F is a piecewise linear function of a and b. So the minimum value occurs at a vertex of the graph. This suggests that we take a numerical approach to finding the appropriate values of a and

b. We illustrate the ℓ^1 regression line with a simple example. Consider the two-variable data $\{(1,1),(2,2),(3,3),(4,5)\}$. The regression line for these data (in the standard metric) is $y = 1.3x - 0.5$. To find the ℓ^1 regression line we need to find the values of a and b that minimize the function

$$F(a,b) = |1-(a+b)| + |2-(2a+b)| + |3-(3a+b)| + |5-(4a+b)|\,.$$

Using numerical methods we find that the minimum value of F is not unique. It is achieved whenever (a,b) is in the triangular region in $\mathbb{R}^2$ with vertices $(1.5,-1)$, $(1.3,-0.3)$, and $(1,0)$.

4. Quantile Regression

For the regression line and the median regression line, deviations of the data above or below the line are measured by a metric. But for many real-world data we may want to treat data points above or below a particular quantile differently. For example, in studies where the dependent variable is poverty (as measured by income level) the lower quartile of the dependent variable is more relevant. For studies on pollution, the upper decile of pollution levels (the dependent variable) is more important because these levels pose a much more significant health risk. In each case a researcher may choose to lower or pull up the line of best fit towards an extreme quantile of the distribution function of the dependent variable.

This suggests that we weight data points above the proposed regression line differently than those below. The quasimetric λ_p provides the needed tool. That is, we can seek a quantile regression line using the quasimetric defined in (4). Indeed, this type of regression, called *quantile regression*, is used extensively in data analysis [1]. In our setting, the quantile regression line (for the p-quantile) is simply the linear function $y = ax + b$ which minimizes the distance $d(\mathbf{y},\mathbf{L})$ between the data vector $\mathbf{y}$ and the linear vector $\mathbf{L}$ defined in (7), in the quasimetric (4). This distance is the function of the two variables a and b given by

$$F(a,b) = \sum_{i=1}^{N} \lambda_p(y_i - (ax_i + b)) = \sum_{i=1}^{N} \lambda_p(y_i - \widehat{y}_i). \tag{8}$$

So the quantile regression line (for the p-quantile) is the line that minimizes the sum of the p-tilted absolute values λ_p of the residuals $y_i - \widehat{y}_i$. Form the definition of λ_p in (3) we have

$$\lambda_p(y_i - \widehat{y}_i) = \begin{cases} (1-p)\,|y_i - \widehat{y}_i|\,, & \text{if} \quad y_i - \widehat{y}_i < 0; \\ p\,|y_i - \widehat{y}_i|\,, & \text{if} \quad y_i - \widehat{y}_i \geq 0. \end{cases}$$

Thus the quantile regression line is the line that minimizes the sum of the weighted absolute values of the residuals $y_i - \widehat{y}_i$, where the weights are $1-p$ for negative residuals and p for positive residuals. For example, if $p = 0.25$

then for $y_i = 3$ and $\widehat{y}_i = 7$ we have $\lambda_p(y_i - \widehat{y}_i) = (0.75)\,|3-7| = 3$; but for $y_i = 7$ and $\widehat{y}_i = 3$ we have $\lambda_p(y_i - \widehat{y}_i) = (0.25)\,|7-3| = 1$. In Figure 2, distances to points above the line are given a weight 0.25 and those below the line are given a weight 0.75. Because of this imbalance, the line that minimizes the sum of these distances would be pulled down towards the lower data points (the first quartile). If $p = 0.99$ the quantile regression line would be pulled up to the 99th percentile of the distribution of the dependent variable, since positive residuals $y_i - \widehat{y}_i$ are now weighted by 0.99.

Quantile regression gives the researcher the choice to distinguish, or weight, either of the extreme quantiles of a population in constructing a linear model (without ignoring the remaining data). So it is particularly useful in situations where extremes of the data are of interest. For this reason quantile regression has been used in studies involving educational attainment and wage inequality, household income and food expenditures, birth weight and prenatal care, happiness and income, and others [1]. For example, welfare rules apply to families with income below the poverty line. If 11 of the population is below the poverty line, then welfare researchers may find the 0.11 income quantile (and lower quantiles) more pertinent to their research. Finding a regression line that targets this quantile of income levels corresponds to using the quasimetric λ_p with $p = 0.11$ to find a quantile regression line.

Calculating the coefficients in quantile regression is no easy matter. In fact, the idea of quantile regression was considered as early as the nineteenth century, but its use only became feasible with the availability of high speed computers and sophisticated numerical methods. The piecewise linear nature of the regression formula (8) also accommodates linear programming techniques, again with the aid of computing devices. Currently, several statistical software programs compute quantile regression; these include SAS, R, and STATA.

5. Conclusion

Many concepts in data analysis can be put in a common setting by stating them in the context of metric spaces. The median, as well as different quantiles of data, can be realized as minimization problems with respect to a metric or quasimetric. Similarly, regression and quantile regression can be viewed as identical metric space problems only with different metrics or quasimetrics.

Acknowledgment

We are grateful to the referee for crucial insights that contributed to clarifying the presentation in this article.

References

[1] L. Hao and D. Q. Naiman, *Quantile Regression*, Sage Publications, Thousand Oaks, CA, 2007.
[2] R. Koenker and K. F. Hallock, *On Quantile Regression*, Journal of Economic Perspectives, **15** (2001), 143–156.
[3] E. Kreyszig, *Introductory Functional Analysis with Applications*, Wiley, New York, 1978.
[4] F. Newberger, A. Safer, and S. Watson, *What is standard about the standard deviation?*, Missouri J. Math. Scienc., **22** (2010), 86–90.
[5] M. F. Triola, *Elementary Statistics*, Pearson, New York, 2007.
[6] W. A. Wilson, *On Quasimetric Spaces*, Amer. J. Math., **53.3** (1931), 675–684.

MSC(2010): 62J05, 62J99

Department of Mathematics and Statistics, California State University-Long Beach, Long Beach, CA 90840
E-mail address: asafer@csulb.edu

Department of Mathematics and Statistics, California State University-Long Beach, Long Beach, CA 90840
E-mail address: ksuaray@csulb.edu

Department of Mathematics and Statistics, California State University-Long Beach, Long Beach, CA 90840
E-mail address: saleem@csulb.edu