Journal of Probability and Statistical Science
24(2), 1-15Aug. 2026

Some Results about Orthogonal Least-Squares Lines

Steven J. Kilner
Department of Mathematics
Monroe Community College, USA

David L. Farnsworth
School of Mathematics and Statistics
Rochester Institute of Technology, USA

ABSTRACT

A set of quantitative bivariate data is augmented with one additional point. Orthogonal least-squares fitting is used for the original set and the augmented set. The possible locations of the two lines, the points of means of the two sets, and the augmenting point are examined.

***Keywords*:** Orthogonal regression; Selected observation; Location of an observation; Location of a fitted line; Location of the point of means; Principal component analysis

1. Introduction

A set consisting of paired data is supplemented with one additional point and orthogonal least-squares regression lines are found for each set. Alternatively, one pair could be selected in a bivariate data set, and fitted lines for the reduced set and for the full set determined. Having a set of $n-1$ unordered points and adding one more point is the same as selecting one of n unordered points to designate as the n^{th} point. The intention is that there is a set of $n-1$ points and one other point.

Some results presented here offer outcomes that are not possible to achieve by the addition of a single point to a set. Those properties provide limitations on the expected outcome of the fitting. For example, it is shown that if the additional point is on either line, then the two lines coincide. These results supply ways for examining the correctness of calculations and graphical displays. Potentially, the properties could impact statistical analyses by imposing boundaries on values or locations.

For orthogonal least-sum-of-squares linear regression or, simply said, orthogonal regression, the sum of the squared perpendicular distances to the fitted line is minimized over all possible y-intercepts and slopes for the fitted line. It is sometimes used to model situations in which both of the paired variables are measured with the same device, thus all measurements have the same error distribution caused by the imprecision in that device's scale. Sometimes, the devices for the two measurements are different but have the same error in variable distribution. If the two devices or instruments have different variability, the model might be Deming regression, in which distances are taken at an angle different from ninety degrees, which reflects the difference, or often more properly the ratio, of the two measurements' variances [2, 10].

Orthogonal regression is used in principal component analysis, where the line provides the major axis of the scatter of points [4, 5, 8].

Unlike simple linear regression, coordinate transformations present a problem for orthogonal regression. A transformation before the fitting yields a line for the transformed data in the new units, but that model might have little or no relevance to the data in the original units [8]. This is caused by general coordinate transformations altering perpendicularity. Rigid translations and rotations do not alter distances or angles, so they might be freely used. Some techniques in principal component analysis are equivalent to converting each data entry in the pair into a z-score,

- Received August 27, 2026, in final form August 2026.
- David L. Farnsworth (corresponding author) is affiliated with School of Mathematics and Statistics, Rochester Institute of Technology, USA
  Email: dlfsma@rit.edu

which irreversibly alters the fit. There continues to be debate about that issue [1, 3, pp. 93–94, 4, pp. 64–65, 5].

2. Notation

In order to establish the notation, consider the set of $n-1$ x,y-pairs

$$S_{(n-1)} = \{(x_i, y_i) | i = 1, 2, \dots, n-1\}.$$

The subscript $(n-1)$ indicates sample size. The set's point of means is

$$M_{(n-1)} = \left(\bar{x}_{(n-1)}, \bar{y}_{(n-1)}\right).$$

The orthogonal least-sum-of-squares line for this set is determined by minimizing the sum of the squares of the perpendicular distances from each pair to the line

$$L_{(n-1)}: y = a_{(n-1)} + b_{(n-1)}x. \tag{1}$$

That sum of squares is

$$SS_{(n-1)} = \sum_{i=1}^{n-1} \frac{\left(y_i - b_{(n-1)}x_i - a_{(n-1)}\right)^2}{b_{(n-1)}^2 + 1}. \tag{2}$$

Setting $\partial / \partial a_{(n-1)}\left(SS_{(n-1)}\right)$ equal to zero gives

$$a_{(n-1)} = \bar{y}_{(n-1)} - b_{(n-1)}\bar{x}_{(n-1)} \tag{3}$$

[9]. Equation (3) shows that line (1) contains the point of means $M_{(n-1)}$. Setting $\partial / \partial b_{(n-1)}\left(SS_{(n-1)}\right)$ equal to zero and using equation (3) to eliminate $a_{(n-1)}$ gives the quadratic equation for the slope $b_{(n-1)}$,

$$\begin{aligned} &\sum_{i=1}^{n-1}\left(\left(x_i - \bar{x}_{(n-1)}\right)\left(y_i - \bar{y}_{(n-1)}\right)\right) b_{(n-1)}^2 \\ &\quad - \left(\sum_{i=1}^{n-1}\left(y_i - \bar{y}_{(n-1)}\right)^2 - \sum_{i=1}^{n-1}\left(x_i - \bar{x}_{(n-1)}\right)^2\right) b_{(n-1)} \\ &\quad - \sum_{i=1}^{n-1}\left(x_i - \bar{x}_{(n-1)}\right)\left(y_i - \bar{y}_{(n-1)}\right) = 0 \end{aligned} \tag{4}$$

[9]. Introduce the notation

$$S_{XX(n-1)} = \sum_{i=1}^{n-1}\left(x_i - \bar{x}_{(n-1)}\right)^2,\ S_{YY(n-1)} = \sum_{i=1}^{n-1}\left(y_i - \bar{y}_{(n-1)}\right)^2,$$
$$\text{and } S_{XY(n-1)} = \sum_{i=1}^{n-1}\left(\left(x_i - \bar{x}_{(n-1)}\right)\left(y_i - \bar{y}_{(n-1)}\right)\right).$$

Equation (4) is

$$S_{XY(n-1)}b_{(n-1)}^2 - \left(S_{YY(n-1)} - S_{XX(n-1)}\right)b_{(n-1)} - S_{XY(n-1)} = 0. \tag{5}$$

The product of the roots is minus one, so there are two perpendicular lines. See Appendix A for the details of the derivations of equations (3) and (5).

When $S_{XY(n-1)} = 0$, equation (5) becomes

$$\left(S_{YY(n-1)} - S_{XX(n-1)}\right)b_{(n-1)} = 0,$$

requiring

$$S_{YY(n-1)} - S_{XX(n-1)} = 0 \text{ or } b_{(n-1)} = 0.$$

If $S_{YY(n-1)} \neq S_{XX(n-1)}$, then $b_{(n-1)} = 0$, so that the line is $y = \bar{y}_{(n-1)}$. Either this horizontal line is the line of best fit or the perpendicular vertical line $x = \bar{x}_{(n-1)}$ is the line of best fit in the orthogonal least squares sense. The identity of the line of best fit is determined by the direction of the larger spread of the data. If $S_{XX(n-1)} > S_{YY(n-1)}$, then the horizontal line is the line of best fit. If $S_{YY(n-1)} > S_{XX(n-1)}$, then the vertical line is the line of best fit. If $S_{XX(n-1)} = S_{YY(n-1)}$, then equation (5) becomes $0 = 0$ and there is no unique line of best fit as any line passing through the point of means $M_{(n-1)}$ results in the same sum of squared orthogonal distances in (2).

If $S_{XY(n-1)} \neq 0$, then equation (8) can be solved for $b_{(n-1)}$ as

$$b_{(n-1)} = \frac{S_{YY(n-1)} - S_{XX(n-1)} \pm \sqrt{\left(S_{YY(n-1)} - S_{XX(n-1)}\right)^2 + 4\left(S_{XY(n-1)}\right)^2}}{2S_{XY(n-1)}}. \quad (6)$$

The plus sign gives the minimum sum of squares in equation (2), i.e., the best fitting line using the squared orthogonal distances, and the minus sign gives the maximum sum of squares, i.e., the worst fitting line using the squared orthogonal distances. See Appendix B for the details. The plus sign means that the slope $b_{(n-1)}$ has the same sign as the correlation coefficient

$$r_{(n-1)} = \frac{S_{XY(n-1)}}{\sqrt{S_{XX(n-1)}\, S_{YY(n-1)}}}$$

for the set $S_{(n-1)}$, which has the same sign as the slope of that set's simple linear regression line. [6, pp. 76–85, 7, p. 284].

An additional point $P: (x_n, y_n)$ is introduced and the set

$$S_{(n)} = S_{(n-1)} \cup \{P: (x_n, y_n)\} = \{(x_i, y_i) | i = 1, 2, \ldots, n\},$$

is formed. This set's point of means and orthogonal least-squares line are

$$M_{(n)} = \left(\bar{x}_{(n)}, \bar{y}_{(n)}\right)$$

and

$$L_{(n)}: y = a_{(n)} + b_{(n)}x,$$

where

$$a_{(n)} = \bar{y}_{(n)} - b_{(n)}\bar{x}_{(n)}$$

and

$$b_{(n)} = \frac{S_{YY(n)} - S_{XX(n)} + \sqrt{\left(S_{YY(n)} - S_{XX(n)}\right)^2 + 4\left(S_{XY(n)}\right)^2}}{2S_{XY(n)}}, \quad (7)$$

with

$$S_{XY(n)} = \sum_{i=1}^{n}\left(\left(x_i - \bar{x}_{(n)}\right)\left(y_i - \bar{y}_{(n)}\right)\right) = S_{XY(n-1)} + \frac{n-1}{n}\left(x_n - \bar{x}_{(n-1)}\right)\left(y_n - \bar{y}_{(n-1)}\right), \quad (8)$$

$$S_{XX(n)} = \sum_{i=1}^{n}\left(x_i - \bar{x}_{(n)}\right)^2 = S_{XX(n-1)} + \frac{n-1}{n}\left(x_n - \bar{x}_{(n-1)}\right)^2, \quad (9)$$

and

$$S_{YY(n)} = \sum_{i=1}^{n}\left(y_i - \bar{y}_{(n)}\right)^2 = S_{YY(n-1)} + \frac{n-1}{n}\left(y_n - \bar{y}_{(n-1)}\right)^2. \quad (10)$$

The formulae in equations (8)-(10) are derived in Appendix C.

In practice, the chance is tiny of obtaining a set of observations such that the orthogonal least-squares regression line is not unique. Example 1 demonstrates some of the behavior of such data.

Example 1: The data set

$$S = \{(a, b), (-a, -b), (b, -a), (-b, a)\}$$

with $a > 0, b > 0$ illustrates the lack of uniqueness of the orthogonal least-squares line for some sets. The origin is the point of means of the set S. The simple least-squares line is $y = 0$, the correlation coefficient $r_{(4)} = 0$, and $S_{XY(4)} = 0$. The orthogonal least-squares line for S is undefined for these four points, because the sum of orthogonal distances to any line through the origin is $2(a^2 + b^2)$. Alternatively, it could be said that any line through the origin is an orthogonal least-squares line for S.

There is every combination of sets and their sets augmented with one point such that either, neither, or both have a unique orthogonal regression line. For example, the set

$$S \cup \{(c, d) | (c, d) \neq (0,0)\}$$

has a unique orthogonal regression line. The set

$$S \setminus \{(a,b)\} = \{(-a,-b),(b,-a),(-b,a)\}$$

has a unique orthogonal regression line. Since any line containing the origin is an orthogonal regression line for the set

$$S \cup \{(0,0)\},$$

neither this augmented set nor S have a unique line.

Because non-uniqueness is so rare, theorems' premises can contain the assumption of uniqueness of all fitted lines met without jeopardizing applicability. A tactic for addressing non-uniqueness is offered following Theorem 8.

3. Relationships between certain lines and points

The first three theorems describe algebraic relationships among the additional point P and the two points of means $M_{(n)}$ and $M_{(n-1)}$. Beginning with Theorem 4, many of the theorems can be applied to any fitting procedure that has the property that the fitted line contains the point of means of the set that determines the line, e.g., simple linear regression. See Xu [10] for other examples of these fitting procedures. The condition that lines $L_{(n-1)}$ and $L_{(n)}$ are unique appears in many of the theorems' premises, but that condition may be relaxed with appropriate considerations, as discussed following Theorem 8.

The relationships

$$\bar{x}_{(n)} - \bar{x}_{(n-1)} = \frac{1}{n}\left(x_n - \bar{x}_{(n-1)}\right) \text{ and } \bar{y}_{(n)} - \bar{y}_{(n-1)} = \frac{1}{n}\left(y_n - \bar{y}_{(n-1)}\right)$$

among the coordinates $\left(\bar{x}_{(n)}, \bar{y}_{(n)}\right)$ of the point of means $M_{(n)}$, the coordinates $\left(\bar{x}_{(n-1)}, \bar{y}_{(n-1)}\right)$ of the point of means $M_{(n-1)}$, and the coordinates (x_n, y_n) of the additional point P are used in the proofs of Theorems 1 and 2. See Appendix C for those relationships.

Theorem 1. The points $P, M_{(n)}$, and $M_{(n-1)}$ coincide or else they are distinct and collinear.

Proof. If $x_n \neq \bar{x}_{(n-1)}$ and $y_n \neq \bar{y}_{(n-1)}$, then the slopes of the lines $M_{(n-1)}P$ and $M_{(n)}P$ are both $\left(y_n - \bar{y}_{(n-1)}\right)/\left(x_n - \bar{x}_{(n-1)}\right)$. Those lines share the point P, so the three points are distinct and collinear. If $x_n \neq \bar{x}_{(n-1)}$ and $y_n = \bar{y}_{(n-1)}$, then the three points are distinct and lie on the line $y = \bar{y}_{(n-1)}$ of slope zero. If $x_n = \bar{x}_{(n-1)}$ and $y_n \neq \bar{y}_{(n-1)}$, then the points are distinct and lie on the vertical line $x = \bar{x}_{(n-1)}$. If $x_n = \bar{x}_{(n-1)}$ and $y_n = \bar{y}_{(n-1)}$, i.e., point P is the point of means $M_{(n-1)}$, then the three points coincide. When the points are collinear, designate the line by L. ■

The inequalities involving sample means in Theorem 2 imply Theorem 3, which says that $M_{(n)}$ must be the central point of the three points on line L that is defined in Theorem 2.

Theorem 2. If $x_n \neq \bar{x}_{(n-1)}$, then the value $\bar{x}_{(n)}$ is strictly between $\bar{x}_{(n-1)}$ and x_n. If $y_n \neq \bar{y}_{(n-1)}$, then $\bar{y}_{(n)}$ is strictly between $\bar{y}_{(n-1)}$ and y_n.

Proof. The six possible orderings of $\bar{x}_{(n)}$, $\bar{x}_{(n-1)}$, *and* x_n are

I. $\bar{x}_{(n-1)} < \bar{x}_{(n)} < x_n$

II. $\bar{x}_{(n-1)} < x_n < \bar{x}_{(n)}$

III. $\bar{x}_{(n)} < \bar{x}_{(n-1)} < x_n$

IV. $\bar{x}_{(n)} < x_n < \bar{x}_{(n-1)}$

V. $x_n < \bar{x}_{(n-1)} < \bar{x}_{(n)}$

VI. $x_n < \bar{x}_{(n)} < \bar{x}_{(n-1)}$.

Only inequalities I and VI can be true. That can be seen by subtracting $\bar{x}_{(n-1)}$ across all of the six inequalities and replacing $\bar{x}_{(n)} - \bar{x}_{(n-1)}$ with $\left(x_n - \bar{x}_{(n-1)}\right)/n$, which give

I. $0 < \left(x_n - \bar{x}_{(n-1)}\right)/n < x_n - \bar{x}_{(n-1)}$

II. $0 < x_n - \bar{x}_{(n-1)} < \left(x_n - \bar{x}_{(n-1)}\right)/n$

III. $\left(x_n - \bar{x}_{(n-1)}\right)/n < 0 < x_n - \bar{x}_{(n-1)}$

IV. $\left(x_n - \bar{x}_{(n-1)}\right)/n < x_n - \bar{x}_{(n-1)} < 0$

V. $x_n - \bar{x}_{(n-1)} < 0 < \left(x_n - \bar{x}_{(n-1)}\right)/n$

VI. $x_n - \bar{x}_{(n-1)} < \left(x_n - \bar{x}_{(n-1)}\right)/n < 0.$

Inequality II is incorrect because it says that a positive number is less than $1/n$ times itself. Inequalities III and V are incorrect because they each say that $x_n - \bar{x}_{(n-1)}$ is both positive and negative. Inequality IV is incorrect because it says that a negative number is more than $1/n$ times itself. ■

Theorem 3. If the point P is not the point of means $M_{(n-1)}$, then the point of means $M_{(n)}$ is strictly between points $M_{(n-1)}$ and P on line L.

Theorem 4. If the points $P, M_{(n)}$, and $M_{(n-1)}$ do not coincide, if lines $L_{(n-1)}$ and $L_{(n)}$ are unique, and if the point P is not on line $L_{(n-1)}$, then line $L_{(n)}$ is not the same as line $L_{(n-1)}$.

Proof. Because the point P is not on the line $L_{(n-1)}$, neither is the point of means $M_{(n)}$ on the line $L_{(n-1)}$. Indeed, if point $M_{(n)}$ were on line $L_{(n-1)}$, then line $L_{(n-1)}$ must contain point P, because the line $L_{(n-1)}$ also contains the point of means $M_{(n-1)}$ and $M_{(n-1)}, M_{(n)}$, and P are distinct and collinear by Theorem 1. That is a contradiction of the premise that the point P is not on line $L_{(n-1)}$.

Line $L_{(n-1)}$ cannot be the line $L_{(n)}$, because the point of means $M_{(n)}$ of $S_{(n)}$ is not on the line $L_{(n-1)}$ and the orthogonal least-squares line $L_{(n)}$ contains $M_{(n)}$. ■

The proof of Theorem 5 is the same as the proof of Theorem 4 *mutatis mutandis*.

Theorem 5. If the points $P, M_{(n)}$, and $M_{(n-1)}$ do not coincide, if lines $L_{(n-1)}$ and $L_{(n)}$ are unique, and if the point P is not on line $L_{(n)}$, then line $L_{(n)}$ is not the same as line $L_{(n-1)}$.

Theorem 6 is a contrapositive to Theorems 4 and 5.

Theorem 6. If the points $P, M_{(n)}$, and $M_{(n-1)}$ do not coincide, if lines $L_{(n-1)}$ and $L_{(n)}$ are unique, and if line $L_{(n)}$ is the same as line $L_{(n-1)}$, then the point P is on line $L_{(n)}$.

The proof of Theorem 7 uses Theorem 1.

Theorem 7. If the points $P, M_{(n)}$, and $M_{(n-1)}$ do not coincide, if lines $L_{(n-1)}$ and $L_{(n)}$ are unique, and if the point P is not on line $L_{(n)}$ nor on line $L_{(n-1)}$, then neither the point of means $M_{(n)}$ nor the point of means $M_{(n-1)}$ can be the intersection point of the two fitted lines $L_{(n)}$ and $L_{(n-1)}$.

Proof. The point of means $M_{(n-1)}$ is on line $L_{(n-1)}$. However, it is seen as follows that the point $M_{(n-1)}$ cannot be on line $L_{(n)}$. Point $M_{(n)}$ is on line $L_{(n)}$. If point $M_{(n-1)}$ were also on line $L_{(n)}$, then, by Theorem 1, point P would be on line $L_{(n)}$, as well, which is a contradiction.

Similarly, the point of means $M_{(n)}$ is on line $L_{(n)}$. However, the point $M_{(n)}$ cannot be on line $L_{(n-1)}$. Point $M_{(n-1)}$ is on line $L_{(n-1)}$. If point $M_{(n)}$ were also on line $L_{(n-1)}$, then, by Theorem 1, point P would be on line $L_{(n-1)}$, as well, which is a contradiction. ■

Theorem 8. If the additional observation P is on the line $L_{(n-1)}$ and if lines $L_{(n-1)}$ and $L_{(n)}$ are unique, then the lines $L_{(n-1)}$ and $L_{(n)}$ coincide and point P is also on line $L_{(n)}$.

Proof. The value of the sum of squares of the perpendicular distances for $S_{(n)}$, which has the selected point P on line $L_{(n-1)}$, is the same as the value associated with $L_{(n-1)}$ by the addition of point P to $S_{(n-1)}$. Any possible line for $L_{(n)}$ that is different from $L_{(n-1)}$ would involve a larger value of the sum of squares for the outcome of the minimization procedure, as follows. The line $L_{(n-1)}$ had been created as the unique fitted line for $S_{(n-1)}$ with a minimum value. For that replacement line for $L_{(n-1)}$, there could be the addition of a positive value contributed by the selected point P, thereby increasing the total. Hence, the line $L_{(n)}$ is the same as the line $L_{(n-1)}$. Since point P is on line $L_{(n-1)}$, P is also on line $L_{(n)}$. ■

The assumption that the set $S_{(n)}$ has a unique orthogonal regression line is not required in Theorem 8, if the line $L_{(n-1)}$ is identified by the location of point P.

Assume that the fitted line for $S_{(n-1)}$ is not unique. All fitted lines for $S_{(n-1)}$ contain the point of means $M_{(n-1)}$ of $S_{(n-1)}$, so each line in the set of fitted lines is distinguished by its slope. Select one of the fitted lines for $S_{(n-1)}$ and keep that choice throughout, naming it $L_{(n-1)}$ and placing point P, which is not the point of means $M_{(n-1)}$, on that line. Alternatively, choose the line $L_{(n-1)}$ from among the fitted lines because it contains the point of interest P. Because the selected point P is on line $L_{(n-1)}$, the value of the sum for $S_{(n)}$ is the same as the value of the sum associated with $L_{(n-1)}$ by the addition of this point P to $S_{(n-1)}$. Any possible line that is the line $L_{(n)}$, which is different from the selected $L_{(n-1)}$, would involve an equal or larger value of the sum of squares than the value for $L_{(n-1)}$ for the outcome of the minimization procedure for the following reasons. The line $L_{(n-1)}$ had been created as a fitted line for $S_{(n-1)}$ with the minimum value for the set $S_{(n-1)}$. So, the contribution by those $n-1$ points to the sum of squares for any other line would be equal to or larger than the value for $S_{(n-1)}$. For any other line $L_{(n)}$, there would be the addition of a positive value to the sum contributed by the selected point P, if point P is not on the line $L_{(n)}$. Line $L_{(n)}$ is any line that had been a candidate for the selected line $L_{(n-1)}$ that also contains the point P. However, the selected $L_{(n-1)}$ is the unique fitted line for $S_{(n-1)}$ that contains P. Hence, the line $L_{(n)}$ is the same as the selected line $L_{(n-1)}$,

The potential non-uniqueness of the fitted line for $S_{(n)}$ is not germane here. Because the line is constrained to pass through both the point P and the point of means, it is uniquely determined.

The proof of the following theorem appears in Appendix D.

Theorem 9. If the points $P, M_{(n)}$, and $M_{(n-1)}$ do not coincide, if the additional point P is on line K that is perpendicular to $L_{(n-1)}$ at the point of means $M_{(n-1)}$, and if lines $L_{(n-1)}$ and $L_{(n)}$ are unique, then lines $L_{(n-1)}$ and $L_{(n)}$ are parallel. These points on line K are the only additional points that yield parallel lines.

Theorem 10. If the points $P, M_{(n)}$, and $M_{(n-1)}$ do not coincide, if lines $L_{(n-1)}$ and $L_{(n)}$ are unique, and if the chosen point P is not on line $L_{(n-1)}$ nor on line $L_{(n)}$, then points P and $M_{(n-1)}$ lie on opposite sides of the line $L_{(n)}$.

Proof. By Theorems 1 and 3, the points $P, M_{(n)}$, and $M_{(n-1)}$ are collinear on L with $M_{(n)}$ strictly between the other two points. Because P is not on lines $L_{(n-1)}$ and $L_{(n)}$, lines L, $L_{(n-1)}$, and $L_{(n)}$ are distinct. Since $M_{(n)}$ is on the line $L_{(n)}$, P and $M_{(n-1)}$ are on opposite sides of line $L_{(n)}$. ■

If the points $P, M_{(n)}$, and $M_{(n-1)}$ do not coincide, if lines $L_{(n-1)}$ and $L_{(n)}$ are unique, and if the chosen point P is not on line $L_{(n-1)}$ nor on line $L_{(n)}$, then it can be said that point P is not between the lines $L_{(n-1)}$ and $L_{(n)}$ and is on the side of line $L_{(n)}$ in the following sense.

If point P is on line K, then by Theorem 9 the lines $L_{(n-1)}$ and $L_{(n)}$ are parallel. Point $M_{(n-1)}$ is on line $L_{(n-1)}$, and point $M_{(n)}$ is on line $L_{(n)}$. By Theorem 3, point $M_{(n)}$ is strictly between points $M_{(n-1)}$ and P.

If point P is not on line K, then by Theorem 9 the lines $L_{(n-1)}$ and $L_{(n)}$ are not parallel. By Theorem 7, those lines' point of intersection is not either point of means. By Theorem 10, points P and $M_{(n-1)}$ lie on opposite sides of the line $L_{(n)}$.

4. Closing comments

This paper examines algebraic and geometric relationships that arise when a set of bivariate data is augmented by a single additional observation in the context of orthogonal least-squares linear fitting. Theorems 1 through 3 establish some underlying geometry: The additional point, the point of means of the original set, and the point of means of the augmented set either coincide or are collinear, with the augmented set's point of means always located strictly between the other two points. This relationship is independent of the fitting procedure.

Theorems 4 through 8 characterize the conditions under which the fitted lines for the original and augmented sets coincide or differ. Notably, placing the additional point on the original fitted line forces the two lines to coincide. Theorems 9 and 10 further restrict the spatial outcomes by defining the unique set of possible positions for the additional point that yield parallel fitted lines, and by showing that the additional point and the original point of means must lie on opposite sides of the line for the augmented data set. The addition of a single point cannot produce arbitrary changes to the fitted line or the point of means; the possible outcomes are strictly limited to a definable set of configurations.

Theorems 4 through 8 are nominally about orthogonal least-squares fitting, but they can be extended beyond orthogonal regression to any fitting procedure that has the property that the point of means is on the fitted line.

The constraints given in the theorems offer practical utility as diagnostic tools. Because they hold for every valid dataset, excluding the rare cases of non-uniqueness as discussed in Sections 2 and 3, they provide a mathematically rigorous check on statistical computations and graphical displays. If a report containing information about points of means fails to correctly display the collinearity guaranteed by Theorem 1, or fails to place the point of means for the augmented set strictly between the original set's point of means and the additional point, as required by Theorem 3, that failure signals a computational or plotting error.

Consequently, these theorems provide valuable limitations on expected outcomes, proving useful wherever datasets are amended incrementally, such as in exploratory data analysis, sequential data collection, and principal component updates.

Acknowledgement

The authors are grateful for the helpful comments on this paper given by Joshua M. Fitzhugh of Midlothian, Virginia.

Conflict of interest and funding

The authors declare that there is no conflict of interest for this paper. No funds were received by the authors regarding this paper.

References

[1]. Abdi, H., and Williams, L. J. (2010). Principal Component Analysis. *WIREs Computational Statistics*, 2(4, July/August), 433–459.
[2]. Deming, W. E. (1964). *Statistical Adjustment of Data*. New York, NY, USA: Dover Publications.
[3]. Gwet, K. L. (2020). *Beginner's Guide to Principal Components*. Gaithersburg, MD, USA: AgreeStat Analytics.
[4]. Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). New York, NY, USA: Springer-Verlag.
[5]. Jolliffe, I. T., and Cadima, J. (2016). Principal Component Analysis: A Review and Recent Developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(April)(No. 2065), 1–16.
[6]. Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). New York, NY, USA: McGraw-Hill/Irwin.
[7]. Montgomery, D. C., Runger, G. C. and Hubele, N. F. (2011). *Engineering Statistics* (5th ed.). New York, NY, USA: Wiley.
[8]. Tofallis, C. (2023). Fitting an Equation to Data Impartially. *Mathematics*, 11(article 3957).
[9]. Weisstein, E. W. (September 16, 2002). Least Squares Fitting –Perpendicular Offsets. Wolfram MathWorld, https://mathworld.wolfram.com/LeastSquaresFittingPerpendicularOffsets.html (accessed August 25, 2026).
[10]. Xu, S. (2014). A Property of Geometric Mean Regression. *The American Statistician*, 68(4, November), 277–281.

Appendix A: The y-intercept and slope of fitted lines

In order to obtain equation (3) for $a_{(n-1)}$, from equation (2), set

$$\frac{\partial}{\partial a_{(n-1)}} SS_{(n-1)} = \frac{\partial}{\partial a_{(n-1)}} \sum_{i=1}^{n-1} \frac{\left(y_i - b_{(n-1)}x_i - a_{(n-1)}\right)^2}{b_{(n-1)}^2+1}$$
$$= \frac{-2}{b_{(n-1)}^2+1} \sum_{i=1}^{n-1} \left(y_i - b_{(n-1)}x_i - a_{(n-1)}\right)$$
$$= \frac{-2}{b_{(n-1)}^2+1} \left(\sum_{i=1}^{n-1} y_i - b_{(n-1)} \sum_{i=1}^{n-1} x_i - (n-1)a_{(n-1)}\right)$$

equal to zero.

In order to obtain equation (5) for $b_{(n-1)}$, set

$$\frac{\partial}{\partial b_{(n-1)}} SS_{(n-1)} = \frac{\partial}{\partial b_{(n-1)}} \sum_{i=1}^{n-1} \frac{\left(y_i - b_{(n-1)}x_i - a_{(n-1)}\right)^2}{b_{(n-1)}^2+1}$$
$$= \frac{\partial}{\partial b_{(n-1)}} \sum_{i=1}^{n-1} \frac{\left(y_i - b_{(n-1)}x_i - \left(\bar{y}_{(n-1)} - b_{(n-1)}\bar{x}_{(n-1)}\right)\right)^2}{b_{(n-1)}^2+1}$$
$$= \frac{\partial}{\partial b_{(n-1)}} \sum_{i=1}^{n-1} \frac{\left(\left(y_i - \bar{y}_{(n-1)}\right) - b_{(n-1)}\left(x_i - \bar{x}_{(n-1)}\right)\right)^2}{b_{(n-1)}^2+1}$$
$$= \frac{\partial}{\partial b_{(n-1)}} \sum_{i=1}^{n-1} \frac{\left(x_i - \bar{x}_{(n-1)}\right)^2 b_{(n-1)}^2 - 2\left(x_i - \bar{x}_{(n-1)}\right)\left(y_i - \bar{y}_{(n-1)}\right)b_{(n-1)} + \left(y_i - \bar{y}_{(n-1)}\right)^2}{b_{(n-1)}^2+1}$$
$$= \frac{\partial}{\partial b_{(n-1)}} \frac{\sum_{i=1}^{n-1}\left(x_i - \bar{x}_{(n-1)}\right)^2 b_{(n-1)}^2 - 2\sum_{i=1}^{n-1}\left(\left(x_i - \bar{x}_{(n-1)}\right)\left(y_i - \bar{y}_{(n-1)}\right)\right)b_{(n-1)} + \sum_{i=1}^{n-1}\left(y_i - \bar{y}_{(n-1)}\right)^2}{b_{(n-1)}^2+1}$$
$$= \frac{\partial}{\partial b_{(n-1)}} \frac{S_{XX(n-1)}b_{(n-1)}^2 - 2S_{XY(n-1)}b_{(n-1)} + S_{YY(n-1)}}{b_{(n-1)}^2+1}$$
$$= \frac{\left(2S_{XX(n-1)}b_{(n-1)} - 2S_{XY(n-1)}\right)\left(b_{(n-1)}^2+1\right) - \left(S_{XX(n-1)}b_{(n-1)}^2 - 2S_{XY(n-1)}b_{(n-1)} + S_{YY(n-1)}\right)2b_{(n-1)}}{\left(b_{(n-1)}^2+1\right)^2}$$
$$= \frac{S_{XY(n-1)}b_{(n-1)}^2 - \left(S_{YY(n-1)} - S_{XX(n-1)}\right)b_{(n-1)} - S_{XY(n-1)}}{\left(b_{(n-1)}^2+1\right)^2}$$

equal to zero.

Appendix B: The identification of the best and worst fitting lines

From Appendix A,

$$\frac{\partial}{\partial b_{(n-1)}} SS_{(n-1)} = \frac{S_{XY(n-1)}b_{(n-1)}^2 - \left(S_{YY(n-1)} - S_{XX(n-1)}\right)b_{(n-1)} - S_{XY(n-1)}}{\left(b_{(n-1)}^2+1\right)^2}.$$

Then,

$$\frac{\partial^2}{\partial b_{(n-1)}{}^2} SS_{(n-1)} = \frac{1}{\left(b_{(n-1)}^2+1\right)^4}\Big(\left(2S_{XY(n-1)}b_{(n-1)} - \left(S_{YY(n-1)} - S_{XX(n-1)}\right)\right)\left(b_{(n-1)}^2+1\right)^2$$
$$- \left(S_{XY(n-1)}b_{(n-1)}^2 - \left(S_{YY(n-1)} - S_{XX(n-1)}\right)b_{(n-1)} - S_{XY(n-1)}\right)2\left(b_{(n-1)}^2+1\right)\Big).$$

To determine the sign of this second derivative, use equation (5) and substitute for $b_{(n-1)}$ from equation (6) to obtain

$$\frac{\partial^2}{\partial b_{(n-1)}{}^2} SS_{(n-1)}$$

$$= \frac{2S_{XY(n-1)} \frac{S_{YY(n-1)} - S_{XX(n-1)} \pm \sqrt{\left(S_{YY(n-1)} - S_{XX(n-1)}\right)^2 + 4\left(S_{XY(n-1)}\right)^2}}{2S_{XY(n-1)}} - \left(S_{YY(n-1)} - S_{XX(n-1)}\right)}{\left(b_{(n-1)}^2 + 1\right)^2}$$

$$= \frac{\pm\sqrt{\left(S_{YY(n-1)} - S_{XX(n-1)}\right)^2 + 4\left(S_{XY(n-1)}\right)^2}}{\left(b_{(n-1)}^2 + 1\right)^2}.$$

Thus, the minimum of the sum of squares is obtained with the plus sign in the expression for the slope and the maximum is obtained with the minus sign.

Appendix C: Expressions for $S_{XY(n)}$, $S_{XX(n)}$, and $S_{YY(n)}$ in equations (8)–(10)

The point of means $M_{(n)}$: $\left(\bar{x}_{(n)}, \bar{y}_{(n)}\right)$ of the set $S_{(n)}$ can be expressed in terms of the coordinates of the point of means $M_{(n-1)}$: $\left(\bar{x}_{(n-1)}, \bar{y}_{(n-)}\right)$ of the set $S_{(n-1)}$ and the coordinates (x_n, y_n) of the additional point P by

$$\bar{x}_{(n)} = \frac{(n-1)\,\bar{x}_{(n-1)} + x_n}{n} = \bar{x}_{(n-1)} + \frac{1}{n}\left(x_n - \bar{x}_{(n-1)}\right) \text{ and } \bar{y}_{(n)} = \bar{y}_{(n-1)} + \frac{1}{n}\left(y_n - \bar{y}_{(n-1)}\right),$$

For $S_{XY(n)}$ in equation (8),

$$\begin{aligned}
S_{XY(n)} &= \sum_{i=1}^{n}\left(\left(x_i - \bar{x}_{(n)}\right)\left(y_i - \bar{y}_{(n)}\right)\right) \\
&= \sum_{i=1}^{n}\left(\left(x_i - \bar{x}_{(n-1)} - \frac{1}{n}\left(x_n - \bar{x}_{(n-1)}\right)\right)\left(y_i - \bar{y}_{(n-1)} - \frac{1}{n}\left(y_n - \bar{y}_{(n-1)}\right)\right)\right) \\
&= \sum_{i=1}^{n=1}\left(\left(x_i - \bar{x}_{(n-1)}\right)\left(y_i - \bar{y}_{(n-1)}\right)\right) - \frac{1}{n}\left(x_n - \bar{x}_{(n-1)}\right)\sum_{i=1}^{n-1}\left(y_i - \bar{y}_{(n-1)}\right) \\
&\quad - \frac{1}{n}\left(y_n - \bar{y}_{(n-1)}\right)\sum_{i=1}^{n-1}\left(x_i - \bar{x}_{(n-1)}\right) + (n-1)\frac{1}{n^2}\left(x_n - \bar{x}_{(n-1)}\right)\left(y_n - \bar{y}_{(n-1)}\right) \\
&\quad + \left(\frac{n-1}{n}\right)^2\left(x_n - \bar{x}_{(n-1)}\right)\left(y_n - \bar{y}_{(n-1)}\right) \\
&= S_{XY(n-1)} + \frac{n-1}{n}\left(x_n - \bar{x}_{(n-1)}\right)\left(y_n - \bar{y}_{(n-1)}\right).
\end{aligned}$$

Replacing each y-value above with its corresponding x-value leads to equation (9) for $S_{XX(n)}$. Replacing each x-value above with its corresponding y-value leads to equation (10) for $S_{YY(n)}$.

Appendix D: Proof of Theorem 9 about parallel fitted lines

Determine the possible locations of points P: (x_n, y_n) such that the slopes of the lines $L_{(n-1)}$ and $L_{(n)}$ are equal. Using equation (7), equating the slopes gives.

$$b_{(n-1)} = b_{(n)} = \frac{S_{YY(n)} - S_{XX(n)} + \sqrt{\left(S_{YY(n)} - S_{XX(n)}\right)^2 + 4\left(S_{XY(n)}\right)^2}}{2S_{XY(n)}}.$$

Using the expressions in Appendix C and the notation

$$A = \frac{n-1}{n}\left(S_{YY(n-1)} - S_{XX(n-1)}\right),\ B = \frac{n-1}{n}S_{XY(n-1)},\ b = b_{(n-1)},$$

$$s = x_n - \bar{x}_{(n-1)}, \text{ and } t = y_n - \bar{y}_{(n-1)},$$

this is

$$b = \frac{A+t^2-s^2+\sqrt{(A+t^2-s^2)^2+4(B+st)^2}}{2(B+st)}$$

or

$$2b(B+st) - (A+t^2-s^2) = \sqrt{(A+t^2-s^2)^2+4(B+st)^2}.$$

Squaring and cancelling gives

$$(b^2-1)(B+st) = bA + b(t^2-s^2).$$

Observing that $bA = (b^2-1)B$ from $b = \left(A+\sqrt{A^2+4B^2}\right)/(2B)$, this is

$$(b^2-1)B + (b^2-1)st = (b^2-1)B + b(t^2-s^2),$$

which is

$$bt^2 - (b^2-1)st - bs^2 = (t-bs)(bt+s) = 0.$$

The first factor $t - bs = 0$, i.e., $y_n - \overline{y}_{(n-1)} = b_{(n-1)}\left(x_n - \bar{x}_{(n-1)}\right)$, says that point P: (x_n, y_n) is on the line $L_{(n-1)}$, which means that the lines L_n and $L_{(n-1)}$ coincide by Theorem 8. Thus, those lines are not parallel.

The second factor gives $t = -(1/b)s$, i.e., $y_n - \overline{y}_{(n-1)} = \left(-1/b_{(n-1)}\right)\left(x_n - \bar{x}_{(n-1)}\right)$, and the conclusion.