# OBTAINING THE EQUATION OF THE BIVARIATE LEAST SQUARES REGRESSION LINE IN HIGH SCHOOL USING ONLY QUADRATIC FUNCTIONS

Humberto J. Bortolossi and David da Costa Pinho
Fluminense Federal University, Brazil
hjbortol@vm.uff.br

*In this paper, following and adapting the ideas of Casella & Berger (2002) and Niven (1981), we present, using only quadratic functions, a simple derivation of the formulas for the coefficients of the bivariate least squares regression line. Therefore, this approach is very suitable for High School students when Calculus and Linear Algebra are not available (as it is the case of Brazil and other countries). We also present an interactive companion GeoGebra applet to enhance graphically and algebraically the keys ideas.*

## BACKGROUND

Regression analysis is an important technique used in Multivariate Statistics. In High School, however, the least squares regression is often used as a "black box": the student simply enters the data into a software or calculator, and the line of best fit is computed. Perhaps this "black box" approach is practiced because the derivation of the formulas for the coefficients of the linear regression by the least squares criterion is usually done using Differential Calculus (Stewart, 2012) or Linear Algebra (Strang, 2005). Moreover, the lack of explanations for the formulas of the coefficients that be accessible to students is criticized, and many justify the absence of this subject in curricula just because it supposedly requires Calculus or Linear Algebra. Here, we try to show that this is not the case.

## USING QUADRATIC FUNCTIONS TO OBTAIN THE FORMULAS FOR THE COEFFICIENTS OF THE BIVARIATE LEAST SQUARES REGRESSION LINE

To fix the ideas, let's consider a very simple case with only three data points: $A = (1, 1)$, $B = (6, 8)$ and $C = (10,4)$ (Figure 1).

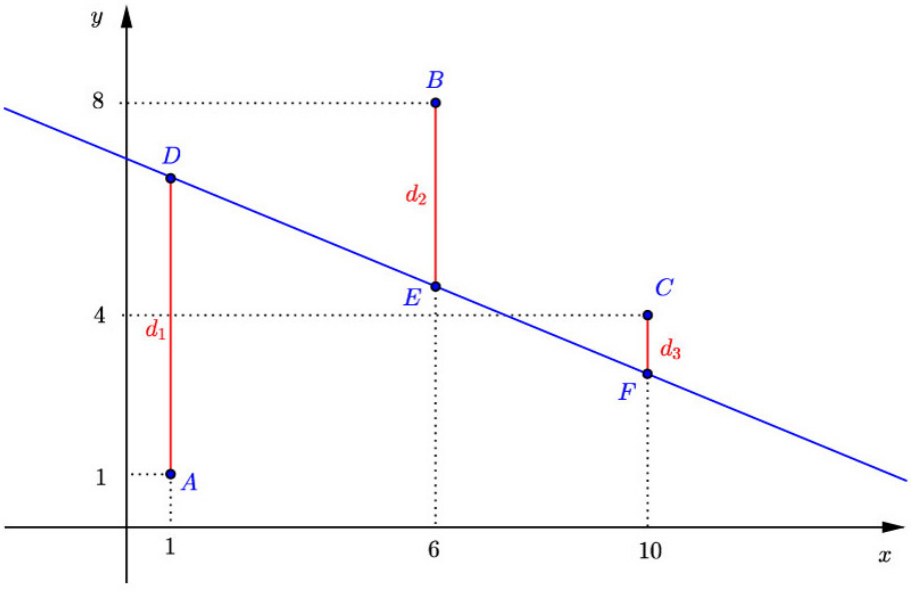


Figure 1. Vertical distances between the points $A = (1, 1)$, $B = (6, 8)$ and $C = (10, 4)$ and the straight line $y = a\,x + b$

We want to find the coefficients $a$ and $b$ of the straight line $y = a\,x + b$ that minimizes the sum of the squares of the vertical distances $d_1$, $d_2$ and $d_3$ between the three points and the sought straight line:

In: M.A. Sorto (Ed.), *Advances in statistics education: developments, experiences and assessments. Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),* July 2015, Rio de Janeiro, Brazil.  iase-web.org/Conference_Proceedings.php

$$\begin{aligned} F(a,b) &= d_1^2 + d_2^2 + d_3^2 = [y_1 - (ax_1 + b)]^2 + [y_2 - (ax_2 + b)]^2 + [y_3 - (ax_3 + b)]^2 \\ &= [1 - (1a + b)]^2 + [8 - (6a + b)]^2 + [4 - (10a + b)]^2 \\ &= [(1 - a) - b]^2 + [(8 - 6a) - b]^2 + [(4 - 10a) - b]^2. \end{aligned}$$

Developing the squares, we obtain that

$$\begin{aligned} F(a,b) &= (1-a)^2 - 2(1-a)b + b^2 + (8-6a)^2 - 2(8-6a)b + b^2 + (4-10a)^2 - 2(4-10a)b + b^2 \\ &= 3b^2 + (34a - 26)b + (1-a)^2 + (8-6a)^2 + (4-10a)^2. \end{aligned}$$

For each value of $a$ (that is, considering $a$ as a parameter), we see that $F(a, b) = d_1^2 + d_2^2 + d_3^2$ is a quadratic function $f$ in the variable $b$:

$$f(b) = 3b^2 + (34a - 26)b + (1-a)^2 + (8-6a)^2 + (4-10a)^2. \quad (1)$$

There is a unique value of $b$ that minimizes this function, namely, the abscissa of the vertex of the parabola that is graph of $f$ (whose formula is widely known in High School):

$$b_v = \frac{-(34a - 26)}{2 \cdot 3} = \frac{-2(17a - 13)}{2 \cdot 3} = \frac{13 - 17a}{3} = \frac{13}{3} - \frac{17a}{3}.$$

Now, replacing $b$ in (1) by $b_v$, it follows that

$$\begin{aligned} F(a,b_v) &= \left[(1-a) - \left(\frac{13}{3} - \frac{17}{3}a\right)\right]^2 + \left[(8-6a) - \left(\frac{13}{3} - \frac{17}{3}a\right)\right]^2 + \left[(4-10a) - \left(\frac{13}{3} - \frac{17}{3}a\right)\right]^2 \\ &= \frac{122}{3}a^2 - \frac{92}{3}a + \frac{74}{3}. \end{aligned}$$

This last expression is a quadratic function $g(a) = F(a,b_v)$ in the variable $a$ whose leading coefficient 122/3 is positive. Thus, this function $g$ admits a unique minimum value at

$$a_v = \frac{-(-92/3)}{2(122/3)} = \frac{23}{61}.$$

Replacing this value in the equation $b_v = (13/3) - (17/3)\ a$, we conclude that $b_v = 134/61$. Therefore, the sought best fit line has the equation $y = (23/61)\ x + 134/61$.

The general case follows the same steps and its study will reveal the structure of the problem. Consider $n$ points in the plane, $(x_1, y_1)$, …, $(x_n, y_n)$, not all having the same abscissa, as shown in Figure 2.

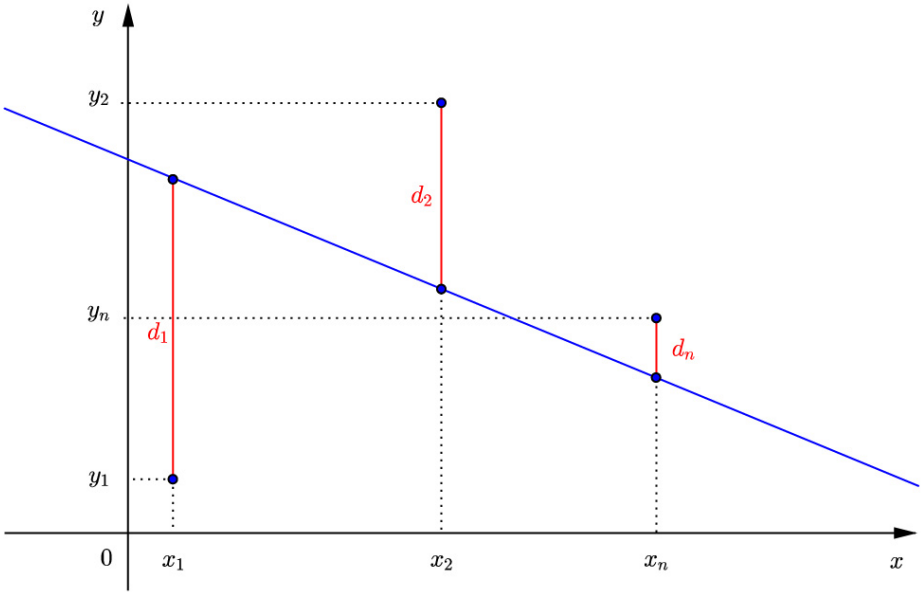


Figure 2. Vertical distances between $(x_1, y_1)$, …, $(x_n, y_n)$ and the straight line $y = a\ x + b$

We want to find the coefficients $a$ and $b$ of the straight line $y = a\ x + b$ that minimizes the sum of the squares of the vertical distances between the $n$ points and the sought straight line:

$$F(a,b) = d_1^2 + \cdots + d_n^2 = (y_1 - (ax_1 + b))^2 + \cdots + (y_n - (ax_n + b))^2$$
$$= ((y_1 - ax_1) - b)^2 + \cdots + ((y_n - ax_n) - b)^2.$$

Developing the squares, we obtain that

$$F(a,b) = d_1^2 + \text{L} + d_n^2 = (y_1 - ax_1)^2 - 2(y_1 - ax_1)\,b + b^2 + \cdots + (y_n - ax_n)^2 - 2(y_n - ax_n)\,b + b^2$$
$$= n\,b^2 - 2[(y_1 - ax_1) + \cdots + (y_n - ax_n)]\,b + (y_1 - ax_1)^2 + \cdots + (y_n - ax_n)^2.$$

For each value of $a$ (that is, considering $a$ as a parameter), we see that $F(a, b) = d_1^2 + d_1^2 + \text{L} + d_n^2$ is a quadratic function $f$ in the variable $b$:

$$f(b) = F(a,b) = n\,b^2 - 2[(y_1 - ax_1) + \cdots + (y_n - ax_n)]b + (y_1 - ax_1)^2 + \cdots + (y_n - ax_n)^2, \qquad (2)$$

where the coefficient of $b^2$ is $n$ and, therefore, it is positive. For this reason, there is a unique value of $b$ that minimizes this function, namely, the abscissa the vertex of the parabola that is graph of $f$:

$$b_v = \frac{-[-2((y_1 - ax_1) + \text{L} + (y_n - ax_n))]}{2 \cdot n} = \frac{(y_1 - ax_1) + \cdots + (y_n - ax_n)}{n} = \frac{y_1 + \cdots + y_n}{n} - a\frac{x_1 + \cdots + x_n}{n}$$
$$= \bar{y} - a\bar{x},$$

where $\bar{x} = (x_1 + \cdots + x_n)/n$ and $\bar{y} = (y_1 + \cdots + y_n)/n$ are, respectively, the means of the variables $x_i$ and $y_i$. Now, replacing $b$ in (2) by $b_v$, it follows that

$$\begin{aligned} F(a,b_v) &= d_1^2 + \cdots + d_n^2 \\ &= [(y_1 - ax_1) - (\bar{y} - a\bar{x})]^2 + \cdots + [(y_n - ax_n) - (\bar{y} - a\bar{x})]^2 \\ &= [(y_1 - \bar{y}) - (x_1 - \bar{x})a]^2 + \cdots + [(y_n - \bar{y}) - (x_n - \bar{x})a]^2 \\ &= (y_1 - \bar{y})^2 - 2a(y_1 - \bar{y})(x_1 - \bar{x}) + (x_1 - \bar{x})^2 a^2 + \cdots + \\ &\quad (y_n - \bar{y})^2 - 2a(y_n - \bar{y})(x_n - \bar{x}) + (x_n - \bar{x})^2 a^2 \\ &= [(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]\,a^2 - 2[(y_1 - \bar{y})(x_1 - \bar{x}) + \cdots + (y_n - \bar{y})(x_n - \bar{x})]\,a + \\ &\quad (y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2. \end{aligned}$$

This last expression is a quadratic function $g(a) = F(a,b_v)$ in the variable $a$ whose leading coefficient $(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2$ is positive once the abscissas $x_1$, …, $x_n$, are not all equal to each other. Thus, this function $g$ admits a unique minimum value at

$$a_v = \frac{-(-2[(y_1 - \bar{y})(x_1 - \bar{x}) + \cdots + (y_n - \bar{y})(x_n - \bar{x})])}{2[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]} = \frac{(y_1 - \bar{y})(x_1 - \bar{x}) + \cdots + (y_n - \bar{y})(x_n - \bar{x})}{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}.$$

Replacing this value in the equation $b_v = \bar{y} - a\bar{x}$, we conclude that

$$b_v = \bar{y} - \frac{(y_1 - \bar{y})(x_1 - \bar{x}) + \cdots + (y_n - \bar{y})(x_n - \bar{x})}{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}\,\bar{x}.$$

So, by the least squares criterion, the sought best fit line $y = a\,x + b$ is that one whose coefficients $a$ and $b$ are given by

$$a = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad b = \bar{y} - \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\bar{x}. \qquad (3)$$

While this general proof may look intimidating, again, in the classroom, the idea is to start with a very simple particular numerical example so that the students may grasp the main ideas.

REMARKS

When Differential Calculus or Linear Algebra (or Physics, as Levi (2009) shows) is used to compute the formulas for the regression line coefficients, different expressions appear:

$$a = \frac{n\sum_{i=1}^{n}x_i y_i - \left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2} \quad \text{and} \quad b = \frac{\left(\sum_{i=1}^{n}x_i^2\right)\left(\sum_{i=1}^{n}y_i\right) - \left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}x_i y_i\right)}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}. \qquad (4)$$

It is possible to show that the two expressions (3) and (4) are equivalent (see, for instance, (Pinho, 2014)). However, the expression (3) has some advantages:

- It is easier to see, using (3), why it is required the hypothesis that the abscissas $x_1$, …, $x_n$ must be not all equal to each other.

- It is easier to see, using (3), why the regression line passes through the point $(\bar{x}, \bar{y})$. In fact, once $y = ax + b$ and $b = \bar{y} - a\bar{x}$, it follows that $y = \bar{y} + a(x - \bar{x})$ and, so, when $x = \bar{x}$, we have $y = \bar{y}$.

The steps described previously for the computation of the coefficients of the line of best fit assumes implicitly that the function $F$ has a global minimum, that is, assuming the existence of a global minimum, the technique of minimizing each coordinate of $F$ indeed obtains an explicit expression for the optimal solution in terms of the input data points. The existence of such (unique) global minimum for $F$ can be also proved without Calculus or Linear Algebra. For the interested reader, we recommend the reference (Niven, 1981, p. 182-183).

## EXPLORING A GEOMETRICAL INTERPRETATION WITH GEOGEBRA

As we have seen, the "line of best fit" $y = a\,x + b$ for a set of points $(x_1, y_1)$, …, $(x_n, y_n)$ (not all having the same abscissa) by the least squares criterion is that one whose coefficients $a$ and $b$ minimize the sum

$$S = F(a,b) = d_1^2 + \cdots + d_n^2,$$

where $d_i$ is the vertical distance from the point $(x_i, y_i)$ to the straight line $y = a\,x + b$. Now, if we build a square whose side is the vertical segment with ends at $(x_i, y_i)$ and $(x_i, a\,x_i + b)$, then the area of this square is equal to $d_i^2$. Thus, building up a square of side $(d_1^2 + \cdots + d_n^2)^{1/2}$, whose area is exactly $S$ (the sum of the areas of each square), we see that to find the "line of best fit" means to find values of $a$ and $b$ that minimize this area $S$. Figure 3 illustrates this idea for $n = 3$, $A = (x_1, y_1) = (1, 1)$, $B = (x_2, y_2) = (6, 8)$ and $C = (x_3, y_3) = (10, 4)$.

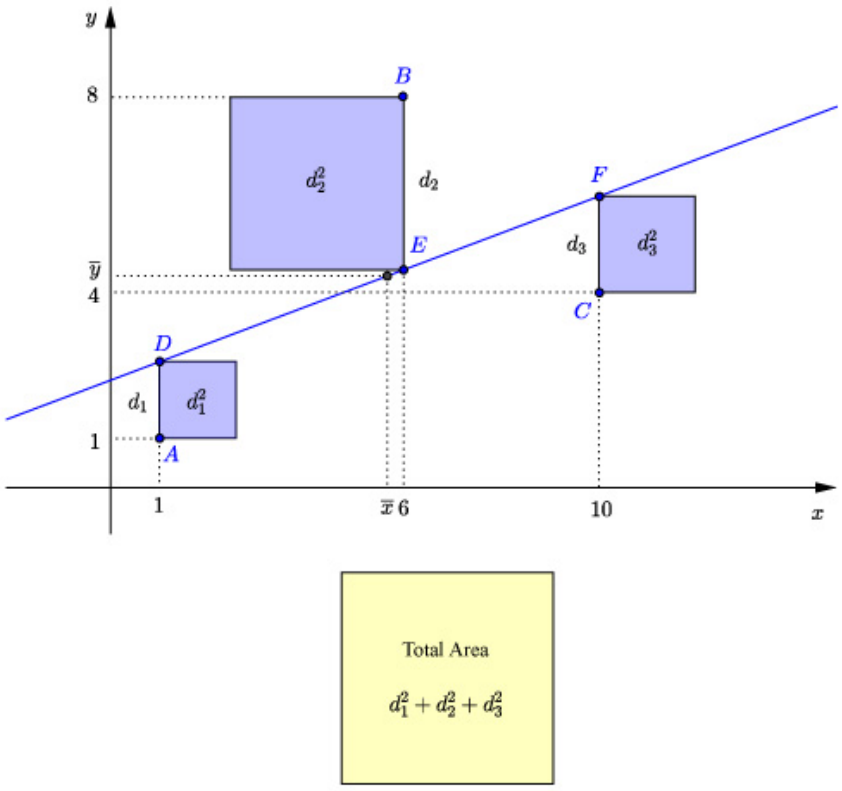


Figure 3. A geometrical interpretation for the bivariate least squares regression method

An interactive version of this geometric interpretation in GeoGebra is available at http://tube.geogebra.org/student/m844479 (Figure 4). In this applet, we can drag and move the data points $A$, $B$, $C$ and, also, the points $P$ and $Q$ that determine the fit line $y = a\,x + b$ (in red). Moving the point $P$ alters the value of the coefficient $b$, while moving the point $Q$ alters the value of the coefficient $a$. Thus, the objective is to choose positions for $P$ and $Q$ such that the total area (the

area of the yellow square) is minimum. Of course, the formulas (2) (or (3)) give the answer directly. However, an alternative approach with students that reinforces the key ideas of the second section is to find the values of $a$ and $b$ interactively, as follows: for a given arbitrary position for $P$ (that is, for a given arbitrary value for the coefficient $b$), adjust the point $Q$ (that is, the value of $a$) such the total area is minimal (for the given value of $b$) and, for this task, the blue graph of the function $S = g(a) = F(a, b)$ displayed on the right of that applet may be handy; done this, choose a new position for $P$ (that is, a new value of $b$) such that the total area is minimal (for the given value of $a$) and, for this task, the red graph of the function $S = f(b) = F(a, b)$ displayed on the right of that applet may be also handy; then repeat the processes alternating between $P$ (the value of $b$) and $Q$ (the value of $a$). Note that, in each step of this process, we are minimizing a quadratic function obtained when $a$ or $b$ is considered fixed, that is, as a parameter. After each interaction, the red line will be closer and closer to the line of best fit (the green line). By the way, this method of minimizing a multivariate function minimizing one variable each time is known in Numerical Analysis as the Aitken Double Sweep Method (Luenberger & Ye, 2008).

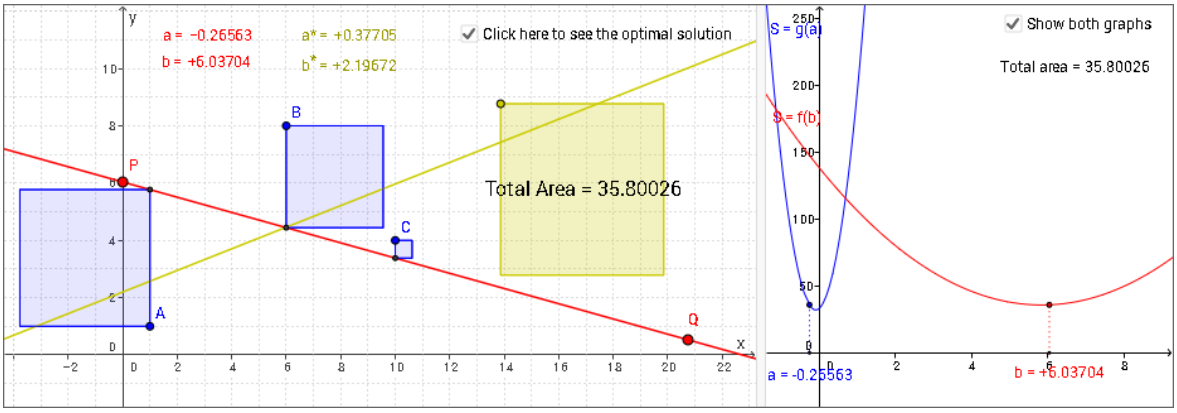


Figure 4. A geometrical interpretation for the bivariate least squares regression method in GeoGebra (http://tube.geogebra.org/student/m844479)

The interactive method described above is slow. There is a slightly different version that is a bit faster: http://tube.geogebra.org/student/m876315. In this applet, as the previous one, if $Q$ is moved and $P$ is kept fixed, only the value of the coefficient $a$ changes. But now, when $P$ is moved and $Q$ is kept fixed, both values of $a$ and $b$ changes, with $a$ given by $a = (y(Q) - y(P))/x(Q) = (y(Q) - b)/x(Q)$. We observe that, yet, we are still minimizing a quadratic function in each interaction.

## FINAL REMARKS

In Brazil, correlation and linear regression are not present in the High School Mathematics Curriculum (Brasil, 2002, 2006), quite probably because of the supposed math involved (curiously, the same document for Biology and Chemistry mentions the use of correlation). In USA, The Principles and Standards for School Mathematics (NCTM, 2000) includes both subjects with the recommendation that "In grades 9–12 all students should […] determine regression coefficients, regression equations, and correlation coefficients using technological tools." (NCTM, 2000), but there are no indications about how to explain the formulas used. In this context, we hope that our work here may facilitate the dissemination of this not so known way to present linear regression since it is more aligned with the High School context (and, besides, it offers an excellent opportunity to see a beautiful application of quadratic functions in Statistics).

In the last two years, the first author has used this approach with in-service teachers in a graduate course on Mathematics Learning and Teaching. When asked if they would explain the linear regression formulas for their students using the approach presented in the second section, they are unanimous: they only would present the case with three points with numerical coordinates once, in their opinion, High School students do not keep attention when generalizations are expounded.

REFERENCES

Brasil. (2002). *PCN+ Ensino Médio: Orientações Educacionais Complementares Aos Parâmetros Curriculares Nacionais. Ciências da Natureza, Matemática e Suas Tecnologias*. Brasília: Ministério da Educação.

Brasil. (2006). *Orientações Curriculares para O Ensino Médio: Ciências da Natureza, Matemática e suas Tecnologias*. Brasília: Ministério da Educação, Brasília.

Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Second Edition. Pacific Grove: Duxbury Thomson Learning.

Levi, M. (2009). *The Mathematical Mechanic: Using Physical Reasoning To Solve Problems*. Princeton: Princeton University Press.

Luenberger, D. G., & Ye, Y. (2008). *Linear and Nonlinear Programming*. Third Edition. New York: Springer-Verlag.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles Standards and for School Mathematics*. Reston: The National Council of Teachers of Mathematics.

Niven, I. (1981). *Maxima and Minima Without Calculus*. Dolciani Mathematical Expositions, Washington, DC: Mathematical Association of America (MAA).

Pinho, D. C. (2014). *Regressão Linear e O Método dos Mínimos Quadrados: Uma Introdução para Professores do Ensino Básico*. Curso de Especialização em Ensino de Matemática, Instituto de Matemática e Estatística, Niterói: Universidade Federal Fluminense.

Stewart, J. (2012). *Calculus*. Seventh Edition. Belmont: Brooks/Cole, Cengage Learning.

Strang, G. (2005). *Linear Algebra and Its Applications*. Fourth Edition. Belmont: Cengage Learning.