Journal of Probability and Statistical Science

20(1), 191-203, August 2022

Estimating the Parameters of a Simple Linear Regression Model Without Using Differential Calculus

Jyotirmoy Sarkar
Indiana University-Purdue University Indianapolis

Mamunur Rashid
DePauw University

ABSTRACT

To estimate the parameters of a simple linear regression model, students who already know calculus can minimize the total squared deviations by setting its first-order partial derivatives to zero and solving simultaneously. For students who do not know calculus, most teachers/textbooks simply state the formulas without justifying them. Students accept the formulas on faith; and for given data, they evaluate the estimates using a calculator or a statistical software. In this paper, we justify the formulas without invoking calculus. We hope the users of statistics will benefit from our proposed justifications.

Keywords: least squares method; partial derivatives; normal equations; positive definite matrix; weighted average

□ Received September 2021, revised December 2021, in final form January 2022.

□ Jyotirmoy Sarkar is affiliated with the Department of Mathematical Sciences, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA and Mamunur Rashid is affiliated with Department of Mathematical Sciences, DePauw University, Greencastle, IN 46135, USA

1. INTRODUCTION

Regression analysis is one of the most used statistical methods (Efron and Tibsirani, 1993). The purpose of regression analysis is to discover a functional relationship (up to error) between a single quantitative response variable (y) and a set of explanatory variables $\{x_1, x_2, \dots, x_p\}$. When the model involves only one quantitative explanatory variable (x) and the functional relationship is linear, it is called a *Simple Linear Regression Model* (SLRM) given by

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

where β_0 and β_1 are respectively the intercept and the slope parameters, and ε is the unobservable error random variable with mean 0 and constant variance σ^2 (free of x).

The first step in fitting a SLRM is to draw a scatter plot of the bivariate data $\{(x_i, y_i): i = 1, 2, \dots, n\}$ to verify visually whether a linear relationship is appropriate. See Figure 1. In an ideal situation, most data points should fall within a certain vertical distance from a line, though a few points may differ from the line substantially. The points must not have any discernable gaps; they must not indicate a superposition of two or more groups; nor should they indicate a relationship that is curvilinear. See, Rothman and Ericson (1987). If a linear relationship is deemed appropriate, only then one should obtain the "best" fitted line that passes through the scatter plot. What makes a proposed line the best?

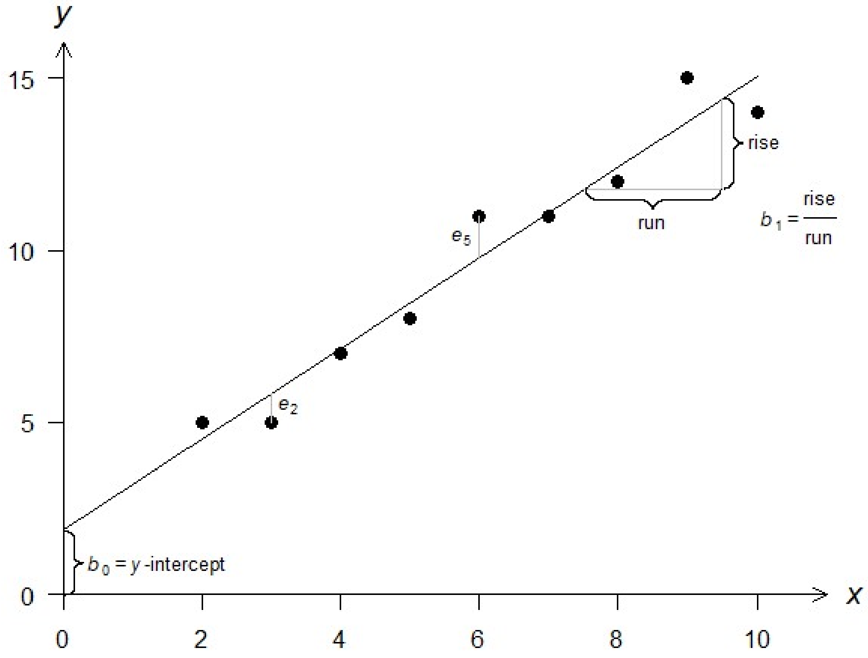

Figure 1. A scatter plot together with a proposed regression line defines the residuals.

For any particular choice of values b_0 of the intercept and b_1 of the slope, one can propose a line $b_0 + b_1$x to obtain fitted values $\hat{y}_i = b_0 + b_1 x_i$ and define the residual as $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$, or the *vertical distance* of any bivariate datum (x_i, y_i) from the fitted value $\hat{y}_i = b_0 + b_1 x_i$. The residual e_i serves as a proxy for the unobservable error ϵ_i. However, since ϵ_i is an unobservable random error component with mean 0, it is natural to require that all residuals sum to 0; that is, $\sum_i e_i = 0$; or equivalently,

$$\bar{y} = b_0 + b_1 \bar{x} \tag{2}$$

Equation (2) is called the *first-moment condition* and is satisfied by any proposed line passing through $(\bar{x}, \bar{y})$, the point where a fulcrum must be placed to balance a physical structure of n identical metal balls placed at $\{(x_i, y_i): i = 1, 2, \dots, n\}$ on a weightless plane. Among the infinitely-many possible such lines passing through $(\bar{x}, \bar{y})$, which one is the best? Another optimization criterion must be invoked.

That second optimization criterion is the *least squares principle*, which was discovered independently by two mathematicians — Carl Friedrich Gauss and Andrien Marie Legendre, resulting in the most famous priority dispute in the history of statistics. Bell (1937, page 227) reports: "When Gauss entered the University of Gottingen in 1795 at the age of eighteen, he had already invented the method of "least squares". However, Legendre published the method in 1806, before Gauss." Stigler (1981) provides evidence that indeed Gauss had used the method in 1799; but it was Legendre who put the method within the reach of the common man. The least squares principle satisfies the *second-moment condition*, or to minimize the sum of squares of all residuals $\sum_i e_i^2$ by choosing b_0 and b_1 appropriately. That is, the problem becomes

$$\min_{b_0, b_1} Q(b_0, b_1) = \min_{b_0, b_1} \sum_{i=1}^{n} \{y_i - (b_0 + b_1 x_i)\}^2 \tag{3}$$

For students who know calculus, solving the minimization problem in (3) is quite straight-forward. Typically, such a student would equate both first-order partial derivatives of $Q(b_0, b_1)$, with respect to b_0 and b_1, respectively, to 0, and solve them simultaneously. Oftentimes, a teacher must remind them to also check that the matrix of second-order derivatives is positive definite. See Binmore and Davies (2001), for example. Indeed, one can check that

$$0 = \frac{\partial Q}{\partial b_0} = \sum_{i=1}^{n} (-2)\{y_i - (b_0 + b_1 x_i)\}$$

$$0 = \frac{\partial Q}{\partial b_1} = \sum_{i=1}^{n} (-2x_i)\{y_i - (b_0 + b_1 x_i)\}$$

whence follow the so-called normal equations

$$\sum_{i=1}^{n} y_i = b_0 n + b_1 \sum_{i=1}^{n} x_i \tag{4}$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} {x_i}^2 \qquad (5)$$

which must be solved simultaneously for b_0 and b_1. Indeed, from (4), we recover the first-moment condition (2). Also, subtracting $\bar{x}$ times (4) from (5), we have

$\sum y_i x_i - n\bar{x}\bar{y} = b_1 \sum (x_i - \bar{x})^2$, or $\sum (y_i - \bar{y})(x_i - \bar{x}) = b_1 \sum (x_i - \bar{x})^2$.

Using short-hand notation

$$S_{xx} = \sum (x_i - \bar{x})^2\,,\ S_{xy} = \sum (y_i - \bar{y})(x_i - \bar{x}) = S_{yx},\ S_{yy} = \sum (y_i - \bar{y})^2 \qquad (6)$$

the solution to the minimization problem in (3) is given by

$$b_1 = \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \text{ and } b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (7)$$

provided that one has verified that the matrix of second-order partial derivatives

$$\begin{bmatrix} \dfrac{\partial^2 Q}{\partial {b_0}^2} & \dfrac{\partial^2 Q}{\partial b_0 \partial b_1} \\ \dfrac{\partial^2 Q}{\partial b_1 \partial b_0} & \dfrac{\partial^2 Q}{\partial {b_1}^2} \end{bmatrix} = 2 \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum {x_i}^2 \end{bmatrix}$$

evaluated at the solution (7), is positive definite by checking that $n > 0, \sum {x_i}^2 > 0$ and the determinant $n \sum {x_i}^2 - (n\bar{x})^2 = n \sum (x_i - \bar{x})^2 > 0$.

How can we establish the slope and the intercept formulas in (7) for students in a non-calculus-based introductory statistics course? Oftentimes, these formulas are given to students without justifying where they come from! See, for example, Johnson and Bhattacharyya (2010), Utts (2015), Larose (2016), McClave and Sincich (2017), Brase and Brase (2018), Navidi and Monk (2019), Peck, Short and Olsen (2019), Kokoska (2020), Veaux, Vellman and Brock (2020). Students are expected to accept the formulas on faith, and then either memorize them or copy them down on a formula sheet.

Thereafter, for any given data set, students are asked to evaluate the slope and the intercept of the “best fitting line” using a calculator or a statistical software and interpret the results.

The purpose of this paper is to estimate the intercept and slope parameters without using calculus so that non-calculus-based students can also derive the formulas, and hence understand and remember them. We shall present two derivations based on high school algebra and coordinate geometry. For ease of presentation, let us begin with some preliminary results on a single variable.

2. MINIMIZING TOTAL DEVIATION OR TOTAL SQUARED DEVIATIONS FOR ONE VARIABLE

Suppose that we want to identify the *center* of a data set on a single variable x. Many different measures are available such as the mean, the median and the mode. If a number A is proposed as the center of a data set, we wish to measure how far “on average” the observations in the data set are from the proposed number A, whether larger or smaller. This can be done either by minimizing (1) the sum of deviations from A given by $L(A) = \sum_{i=1}^{n}|x_i - A|$, or (2) the sum of squared deviations from A given by $Q(A) = \sum_{i=1}^{n}(x_i - A)^2$. We can adopt a proposed A as a good measure of a center of the given data when L(A) is small, or when $Q(A)$ is small.

We leave it to the reader to check that $L(A)$ is minimized when A is a median of the data (that is, after sorting the data, A is the middlemost value if n is odd, or A is any value in between the two middlemost values if n is even). Thus, according to the *least total* *deviation principle*, the best measure of center is the median, and $\mathrm{L}(A)/\mathrm{n}$ is called the mean deviation (MD).

Next, according to the *least squared deviation principle*, one must choose A to minimize $Q(A)$, or

$$\min_{A} \sum_{i=1}^{n} (x_i - A)^2 \qquad (8)$$

Using calculus, one solves the problem by setting $0 = \frac{\partial Q(A)}{\partial A} = \sum_{i=1}^{n}(-2)\{x_i - A\}$, whence $A = \bar{x}$. After checking that $\frac{\partial^2 Q(A)}{\partial A^2}|_{A\,=\,\bar{x}} = 2n > 0$, one has proved that $Q(A)$ is minimized at $A = \bar{x}$.

Students who have not studied calculus can also derive this “best” measure of center using high school algebra as follows: For all A, note that

$$Q(A) = \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - A)^2$$
$$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2(\bar{x} - A)\sum_{i=1}^{n}(x_i - \bar{x}) + n(\bar{x} - A)^2$$
$$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - A)^2 \geq \sum_{i=1}^{n}(x_i - \bar{x})^2 = Q(\bar{x})$$

Where, in the second equality, we have used $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Thus, according to the *least squared deviation principle*, the best measure of center is $A = \bar{x}$ and the associated measure of variation is $Q(\bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2$. Because the $(x_i - \bar{x})$’s satisfy a linear constraint (namely, their sum equals 0), dividing $Q(\bar{x})$ by $(n-1)$, we obtain the “mean” squared deviation (*M*SD). However, *M*SD comes in a square unit (relative to the unit of x_i’s), and hence is difficult to display visually. Therefore, extracting its positive square-root, we get the sample standard deviation (SD) given by

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{S_{xx}}{n-1}} \qquad (9)$$

which is a measure of spread from the best measure of center. Thus, the SD is the root "mean" squared deviation (R*M*SD) from the mean. See, for example, Sarkar and Rashid (2016) and Martin (2003) as to why we divide by $(n-1)$.

The standard deviation not only tells us (on "average") how far the numbers in a data set are from the center (the mean), but also on "average" how far apart the numbers are from one another. Indeed, the sum of differences between all pairs of numbers is

$$\sum_i \sum_{j \neq i}(x_i - x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j) = \sum_{i=1}^{n}\left(nx_i - \sum_{j=1}^{n} x_j\right) = \sum_{i=1}^{n} n(x_i - \bar{x}) = 0$$

and the sum of squares of differences between all possible pairs is

$$\begin{aligned} Q_{xx} &= \sum_i \sum_{j \neq i}(x_i - x_j)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i^2 + x_j^2 - 2x_i x_j) \\ &= \sum_{i=1}^{n}\left(nx_i^2 + \sum_{j=1}^{n} x_j^2 - 2x_i n\bar{x}\right) = 2n\sum_{i=1}^{n} x_i^2 - 2n^2\bar{x}^2 \\ &= 2n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = 2n\, S_{xx} = 2n\,(n-1)s_x \qquad (10) \end{aligned}$$

Since there are $\binom{n}{2}$ possible pairs, the "average" squared difference between any two numbers in the data set is $Q_{xx}/\binom{n}{2} = 4s_{xx}$, and the root mean squared deviation between any two numbers is $\sqrt{4s_{xx}} = 2s_x$, or twice the sample SD of x. See Sarkar and Rashid (2017).

Following similar reasoning, one can simplify the sum of squares of y-differences $Q_{yy} = \sum_i \sum_{j \neq i}(y_i - y_j)^2$ and the sum of products of x-differences with corresponding y-differences $Q_{xy} = \sum_i \sum_{j \neq i}(x_i - x_j)(y_i - y_j)$ to obtain

$$Q_{yy} = 2n\, S_{yy} = 2n\,(n-1)s_{yy},\ Q_{xy} = 2n\, S_{xy} = 2n\,(n-1)s_{xy} \qquad (11)$$

Therefore, the root mean squared deviation between any two y-values is $\sqrt{4s_{yy}} = 2s_y$, and the root mean product of x-differences and y-differences is $\sqrt{4s_{xy}} = 2s_{xy}$.

3. NON-CALCULUS DERIVATIONS OF ESTIMATES OF PARAMETERS IN A SLRM

Method 1: To solve the bivariate minimization problem in (3), first we substitute (2) in (3), so that the objective function reduces to a single variable. That is, since b_0 is a (linear) function of b_1, given by $b_0(\mathrm{b}_1) = \bar{y} - b_1\bar{x}$, we must choose $b_1$to minimize

$$Q(b_0(b_1), b_1) = \sum_{i=1}^{n}\{(y_i - \bar{y}) - b_1(x_i - \bar{x})\}^2 = S_{yy} - 2b_1S_{xy} + {b_1}^2S_{xx}$$

in view of (6). Note that $Q(b_0(b_1), b_1)$ is a quadratic function in the single unknown b_1, which we can rewrite by completing the square as

$$Q(b_0(b_1), b_1) = \left(b_1\sqrt{S_{xx}} - \frac{S_{xy}}{\sqrt{S_{xx}}}\right)^2 + S_{yy} - \frac{{S_{xy}}^2}{S_{xx}} \geq S_{yy} - \frac{{S_{xy}}^2}{S_{xx}} \qquad (12)$$

with equality if and only if $b_1 = S_{xy}/S_{xx}$, the exact same formula as found in (7)!

Method 2: We already reasoned in (2) that the best-fitted line must pass through $(\bar{x}, \bar{y})$. It remains to choose only the slope of the best-fitted line. We propose to choose a weighted average of slopes obtained from each of the $\binom{n}{2}$ possible pairs of points.

For any pair of points (x_i, y_i) and (x_j, y_j) with $i \neq j$, the slope is the quotient $\frac{y_i - y_j}{x_i - x_j}$, provided $x_i \neq x_j$. See Figure 2. Although such a slope is undefined whenever $x_i = x_j$, any concern vanishes if we assign a weight of 0 to such cases. Of course, all weights must be non-negative, which motivates us to propose weights

$$w_{ij} \propto \left(x_i - x_j\right)^2, \quad \text{or} \quad w_{ij} = \frac{\left(x_i - x_j\right)^2}{Q_{xx}}. \tag{13}$$

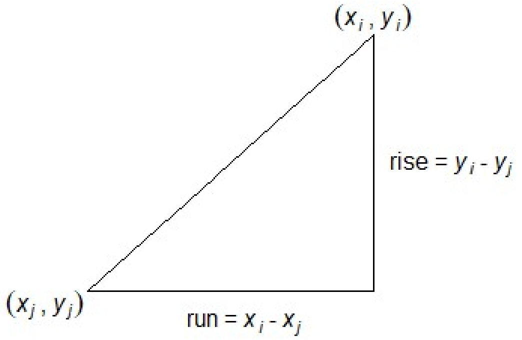

Figure 2. For every pair of points (x_i, y_i) and (x_j, y_j), the ratio of rise over run gives an estimate of slope. How should these estimates be combined into one measure of slope?

A combined measure of slope $\tilde{\beta}_{1,w}$, given by the weighted average of all $\binom{n}{2}$ pairwise slopes with weights $w_{ij} = \left(x_i - x_j\right)^2 / Q_{xx}$, simplifies to

$$\tilde{\beta}_{1,w} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \frac{y_i - y_j}{x_i - x_j} = \frac{1}{Q_{xx}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(x_i - x_j\right)\left(y_i - y_j\right) = \frac{Q_{xy}}{Q_{xx}}$$

In view of (10) and (11), the weighted average of the pairwise slope is

$$\tilde{\beta}_{1,w} = \frac{Q_{xy}}{Q_{xx}} = \frac{S_{xy}}{S_{xx}} \tag{14}$$

which is the same as the least-squares estimate of the slope $\hat{\beta}_1$. Of course, in view of (2), the intercept parameter is estimated as $\tilde{\beta}_{0,w} = \bar{y} - \tilde{\beta}_{1,w}\bar{x} = \bar{y} - \hat{\beta}_1\bar{x} = \hat{\beta}_0$.

Note that Method 2 does not require us to minimize any objective function at all! As such, there is no need to utilize any calculus technique — high school algebra suffices.

4. ESTIMATION OF ERROR SD

We already explained that the sample SD $s_x = \sqrt{\frac{S_{xx}}{n-1}}$ measures the "average" deviation of x-variable from its center $\bar{x}$, as well as half the "average" distance between pairs of x-values. Similarly, the sample SD $s_y = \sqrt{\frac{S_{yy}}{n-1}}$ measures the "average" deviation of y-variable from its center $\bar{y}$, and half the "average" distance between pairs of y-values. Likewise, in view of (12), $\sqrt{\frac{Q(b_0,b_1)}{n-2}} = \sqrt{\frac{S_{yy}-{S_{xy}}^2/S_{xx}}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}}$ measures the "average" deviations among the residuals, and hence estimates the SD of the unobservable error random variables ε_i in the SLRM. The reason for $(n-2)$ in the denominator is that the residuals e_i's satisfy not one, but two linear constraints: $\sum e_i = 0$ and $\sum e_i x_i = 0$.

Finally, of the total variation in random variable Y, measured by S_{yy}, the portion attributed to its linear dependence on x is $\hat{\beta}_1^2\, S_{xx} = {S_{xy}}^2/S_{xx}$. Hence, the proportion of the total variation in Y accounted for by its linear dependence on x is given by

$$R^2 = \frac{{S_{xy}}^2/S_{xx}}{S_{yy}} = \frac{{S_{xy}}^2}{S_{xx}\,S_{yy}} \tag{14}$$

and is called the *coefficient of determination*. In view of Cauchy-Schwarz inequality, see Steele (2004), or simply because (12) is non-negative, we have $0 \le R^2 \le 1$. The higher the R^2, the better the SLRM fits the data. The square root of the coefficient of determination (times the sign of the estimated slope $\hat{\beta}_1$) is called the *correlation coefficient*, which measures the strength and the direction of the simple linear relationship between x an Y.

5. DISCUSSION

We reviewed the calculus-based derivation of the estimates of the parameters in a SLRM and then presented two non-calculus-based derivations. We hope all users of statistics, especially students without the knowledge of calculus, will benefit from these derivations that help them understand and remember better the estimators.

ACKNOWLEDGEMENT

We acknowledge a generous book grant through the Office of Faculty Development at our institution. We thank some inquisitive students who asked for a simpler justification for the regression formulas.

References

[1] Bell, Eric Temple (1937). *Men of Mathematics: The Lives and Achievements of the Great Mathematicians from Zeno to Poincare*, New York, NY: Simon & Schuster.

[2] Binmore, K. and Davies, J. (2007). *Calculus Concepts and Methods*, New York, NY: Cambridge University Press.

[3] Brase, C. H. and Brase, C. P. (2018). *Understandable Statistics: Concepts and Methods* (12th edn), Boston, MA: Cengage Learning.

[4] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Boca Raton, FL: Chapman & Hall/CRC.

[5] Johnson, R. A. and Bhattacharyya, G. K. (2010). *Statistics: Principles and Methods*, New York, NY: Wiley and Sons, Inc.

[6] Kokoska, S. (2020). *Introductory Statistics: A Problem Solving Approach* (3rd edn), New York, NY: W. H. Freeman and Company/MacMillan Learning.

[7] Larose, D. T. (2016). *Discovering Statistics* (3rd edn), New York, NY: W. H. Freeman and Company/MacMillan Learning.

[8] McClave, J. and Sincich, T. (2017). *Statistics* (13th edn), Boston, MA: Pearson.

[9] Martin, M. (2003), "'It's Like... You Know': The Use of Analogies and Heuristics in Teaching Introductory Statistical Methods," *Journal of Statistics Education*, 11(2). Available at http://jse.amstat.org/v11n2/martin.html

[10] Navidi, W. and Monk, B. (2019). *Elementary Statistics* (3rd edn), New York, NY: McGraw-Hill Education.

[11] Peck, R., Short, T., and Olsen, C. (2019). *Introduction to Statistics and Data Analysis* (6th edn), Boston, MA: Cengage Learning.

[12] Rothman, E. and Ericson, W. (1987). *Statistics: Methods and Applications* (2nd ed.), Dubuque, Iowa: Kendall/Hunt Publishing Co.

[13] Sarkar, J. and Rashid, M. (2016). Visualizing the mean, the median, the mean deviation and the standard deviation of a set of numbers, *The American Statistician*, 70(3), 304-312.

[14] Sarkar, J. and Rashid, M. (2017). Visualizing the Sample Standard Deviation, *Educational Research Quarterly*, 40(4), 45-60,

[15] Steele, J. (2004). *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*, New York, NY: Cambridge University Press.

[16] Stigler, S. (1981). Gauss and the Invention of Least Squares, *Annals of Statistics*, 9(3): 465–474. doi:10.1214/aos/1176345451.

[17] Utts, J. (2015). *Seeing Through Statistics* (4th edn), Stanford, CT: Cengage Learning.

[18] Veaux, R. D., Vellman, P., and Brock, D. (2020). *Stats: Data and Models* (5th edn), New Jersey, NJ: Pearson