and so

$$(\sigma_x)^2 < n\sigma_{x^2}.$$

Therefore, $D = f_{mm} f_{bb} - (f_{mb})^2 = 4(n\sigma_{x^2} - (\sigma_x)^2) > 0$, and we are done.

Acknowledgment. I wish to thank the referee for helpful suggestions and comments.

References

1. R. L. Finney, M. D. Weir, and F. R. Giordano, *Thomas' Calculus, Early Transcendentals*, 10th ed., Addison-Wesley Longman, 2001.
2. H. Anton, *Calculus*, 6th ed., Wiley, 1999.

A Painless Approach to Least Squares

Eric S. Key (ericskey@uwm.edu), University of Wisconsin-Milwaukee, Milwaukee, WI 53201

Back in the dark ages of slide rules when I was in high school we had a chemistry lab assignment in which we were instructed to fit a line to the data collected in our experiment. Our sole guideline was to make the line look reasonable. In what was probably my only mathematical inspiration in high school, I figured that one should be able to calculate what this line was if one had had a criteria for "best." Not knowing anything about the least squares criterion, I decided the "best" line should pass through the average point and have slope equal to the average of all line segments passing through at least two data points.

As we know, as far as the least squares criterion goes, I got it half right: the line does pass through the average point. In what follows we will see that that intuition leads to an algebraically simple derivation of the the equation of the line the gives the best fit according to the least squares criterion.

Linear regression and the least squares criterion. The method of least squares for fitting a curve to data was first published by A. M. Legendre in 1805. The problem to be solved is, given a set of functions $\mathcal{F}$ and a set of data points $\{(x_k, y_k), k = 1, 2, \ldots, N\}$, minimize

$$\mathrm{E}[f] := \sum_{k=1}^{N} (f(x_k) - y_k)^2$$

over all $f \in \mathcal{F}$. In other words, find the function f whose graph is "closest" to the data.

For the sake of simplicity, we will consider $\mathcal{F} = \{f(x) = mx + b : -\infty < m < \infty, -\infty < b < \infty\}$. This particular version of the problem is called linear regression, and is what my high school chemistry teacher had in mind. In this case we can think of the problem of one of choosing real numbers m and b to minimize

$$\mathrm{E}[m, b] := \sum_{k=1}^{N} (mx_k + b - y_k)^2$$

over all choices of real numbers m and b, or, if you like, over all slopes and intercepts of non-vertical lines.

There are four approaches to this least squares problem commonly found in textbooks.

The first is to present the results as a black box, either by presenting formulas for finding the regression coefficients or by directing the student to use a graphing calculator or other computing device. For example, this is the approach taken in Moore and McCabe [**4**].

A second approach, usually found in more advanced undergraduate statistics books such as Wackerly, Mendenhall, and Schaeffer [**6**] or in calculus texts such as Simmons [**5**], is to treat $\mathrm{E}[m, b]$ as function of two real variables and by means of differential calculus, deduce its minimum. Indeed, this approach is discussed by Dunn in the preceding Capsule.

Advanced statistics books such as Bickel and Doksum [**1**], and linear algebra texts such as Johnson, Riess and Arnold [**3**], solve the least squares problem using vector projection.

Finally, some authors such as Casella and Berger [**2**] sequentially complete the square in the expression for $\mathrm{E}[m, b]$.

Of all of these methods, only the last is at all suitable for students with limited mathematical experience who want to see for themselves what is going on, and even the last method is rather messy when it comes to using it to produce the standard formulas for the slope and intercept of the "best" line. I will give another method that uses completing the square only once. It is remarkable that it seems to be absent from the literature for the following reason. Almost every author who discusses regression notes that the "average point" mentioned in the introduction lies on the graph of the "best" function, but no one seems to leverage this into the derivation below.

From inspiration to derivation. Suppose that we have $N \geq 2$ data points (x_k, y_k) and that at least two of the x-coordinates of these points are different. We define the average data point $(\overline{x}, \overline{y})$ in the natural way:

$$\overline{x} = \frac{1}{N}\sum_{k=1}^{N} x_k, \qquad \overline{y} = \frac{1}{N}\sum_{k=1}^{N} y_k$$

If we were right that the line of best fit passes through this average point, it would have an equation of the form $y = m(x - \overline{x}) + \overline{y}$. Since this is speculation, we assume that the line has an equation of the form $y = m(x - \overline{x}) + \overline{y} + E$ and proceed to try to determine the values of E and m that minimize the sum of squared errors SS given by

$$SS(m, E) := \sum_{k=1}^{N} \left(m(x_k - \overline{x}) + \overline{y} + E - y_k\right)^2.$$

Let us pause to consider what would happen if we were to multiply out and collect terms in $SS(m, E)$. We see that, if we keep the $(x_k - \overline{x})$'s and $(y_k - \overline{y})$'s unexpanded, we would have expressions of the form

$$\sum_{k=1}^{N}(x_k - \overline{x}), \quad \sum_{k=1}^{N}(y_k - \overline{y}), \quad \sum_{k=1}^{N}(x_k - \overline{x})^2, \quad \sum_{k=1}^{N}(y_k - \overline{y})^2, \quad \text{and}$$

$$\sum_{k=1}^{N}(x_k - \overline{x})(y_k - \overline{y}).$$

For that reason, we introduce the notation

$$S_x^2 = \sum_{k=1}^{N} (x_k - \overline{x})^2. \qquad S_y^2 = \sum_{k=1}^{N} (y_k - \overline{y})^2, \quad \text{and} \quad S_{xy} = \sum_{k=1}^{N} (x_k - \overline{x})(y_k - \overline{y}).$$

Next we observe that

$$\sum_{k=1}^{N} (x_k - \overline{x}) = \sum_{k=1}^{N} (y_k - \overline{y}) = 0.$$

Now, we go ahead and expand our expression for $SS(m, E)$, and with our notation and observation in hand, we see

$$\begin{aligned} SS(m, E) &= \sum_{k=1}^{N} \big(m^2 (x_k - \overline{x})^2 + (y_k - \overline{y})^2 + E^2 \\ &\qquad - 2m(x_k - \overline{x})(y_k - \overline{y}) + 2mE(x_k - \overline{x}) - 2E(y_k - \overline{y})\big) \\ &= S_x^2 m^2 - 2S_{xy} m + S_y^2 + NE^2. \end{aligned}$$

Notice that we may only vary m and E, and that all other expressions are constant. Indeed, our intuition is correct in that to have a minimum we should have $E = 0$. Since

$$SS(m, 0) = S_x^2 m^2 - 2S_{xy} m + S_y^2,$$

our minimization problem is now one of finding the vertex of a parabola, which can easily be solved by completing the square in m. When we do so, we see that the best slope is given by

$$m = \frac{S_{xy}}{S_x^2}.$$

Thus the least-squares-fit line is

$$y = \frac{S_{xy}}{S_x^2} (x - \overline{x}) + \overline{y}.$$

Extensions. One can try to extend this method to the problem of fitting a hyperplane to data in n-dimensional space. Although the average point will still lie on the best hyperplane, we are still left with the problem of determining the normal to the hyperplane, and the solution of this problem by purely algebraic means requires completing the square in more than one variable.

References

1. P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, 1977.
2. G. Casella and R. L. Berger, *Statistical Inference*, Duxbury/Thomson Learning, 2002.
3. L. W. Johnson, R. D. Riess, and J. T. Arnold, *Introduction to Linear Algebra*, Addison-Wesley, 2002.
4. D. S. Moore and G. P. McCabe, *Introduction to the Practice of Statistics*, W. H. Freeman, 1993.
5. G. F. Simmons, *Calculus with Analytic Geometry*, McGraw-Hill, 1996.
6. D. Wackerly, W. Mendenhall II, and R. L. Schaeffer, *Mathematical Statistics with Applications* (6th ed.), Duxbury/Thomson Learning, 2002.