A Quick Proof that the Least Squares Formulas Give a Local Minimum

W. M. Dunn III (willd@nhmccd.edu), Montgomery College, Conroe, TX 77384

A common problem in multivariable calculus is to derive formulas for the slope and y-intercept of the least squares linear regression line, $y = mx + b$, of a given data set of n distinct points, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, with all the x_i different. We recall that the problem is to find the pair (m, b) that minimizes the function

$$w = f(m, b) = \sum_{i=1}^{n} (mx_i + b - y_i)^2,$$

the sum of the squares of the deviations.

Whereas deriving the formulas, most often by setting the gradient of f equal to the zero vector and solving for m and b (see below) is a relatively straightforward computation, using the second partials test to verify that the formulas provide a local minimum is not so easy. The crucial part of the proof is to show that

$$D = f_{mm} f_{bb} - (f_{mb})^2 > 0 \tag{1}$$

holds at the critical pair. ([**1**] proves this by induction, see also [**2**]).

The purpose of this capsule is to show how one can use the Cauchy-Schwartz inequality to give a quick and relatively self-contained proof of inequality (1). We feel that this proof would work well in either lectures or given as a project. It also shows another use of the Cauchy-Schwartz inequality outside of the context of vector algebra, and we hope it becomes more widely known.

For ease of notation, let

$$\sigma_{x^2} = \sum_{i=1}^{n} {x_i}^2, \qquad \sigma_x = \sum_{i=1}^{n} x_i, \qquad \sigma_{xy} = \sum_{i=1}^{n} x_i y_i, \qquad \sigma_y = \sum_{i=1}^{n} y_i.$$

Setting the gradient

$$\nabla f(m, b) = \langle f_m, f_b \rangle = \big\langle 2(m\sigma_{x^2} + b\sigma_x - \sigma_{xy}), 2(m\sigma_x + bn - \sigma_y) \big\rangle$$

equal to the zero vector, and solving for m and b gives the standard formulas [**1**]

$$m = \frac{\sigma_x \sigma_y - n\sigma_{xy}}{(\sigma_x)^2 - n\sigma_{x^2}},$$

and

$$b = \frac{1}{n}(\sigma_y - m\sigma_x).$$

By the second partials test, this pair will be a local minimum if $f_{mm} > 0$ and $D = f_{mm} f_{bb} - (f_{mb})^2 > 0$. Clearly, $f_{mm} = 2\sigma_{x^2} > 0$. It remains to prove $D > 0$. To this end, note that $f_{bb} = 2n$, $f_{mb} = 2\sigma_x$.

Let $\vec{v} = \langle 1, 1, \ldots, 1 \rangle$ and $\vec{w} = \langle x_1, x_2, \ldots, x_n \rangle$. Note that these vectors are not parallel since all the x_i are different. By the Cauchy-Schwartz inequality, $|\vec{v} \cdot \vec{w}| < \|\vec{v}\| \|\vec{w}\|$. Hence, we get

$$|\sigma_x| < \sqrt{n}\sqrt{\sigma_{x^2}},$$

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

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