Derivation of the Least Squares Estimator with Examples

by
Al Bernstein
Signal Science, LLC
www.signalscience.net

Abstract

This paper derives the Least Squares Estimator equations for curve fitting applications and shows two examples using symbolic math. The derivation is interesting because I have not seen a derivation in this form before. In addition, there are interesting math steps that are used – such as taking derivatives of a vector in a matrix equation. Using symbolic examples show the concepts of Least Square curve fitting in a new light.

Least Squares Derivation

The first part of this paper shows the derivation of a weighted least squares optimal estimator for use in curve fitting.

The curve fitting equation is shown in equation (1.1)

$$f(z) = \sum_{l=0}^{M} c_l P_l(z) \tag{1.1}$$

where $f(z)$ is the model that represents a continuous curve that models the data
$P_l(z)$ are the basis set of functions – 0 to M
c_l are the coefficients for the basis functions that provide a best fit to the data using a mean squared error criteria

Digital computers need sampled data – data at discrete data points. The discrete version of (1.1) is given by equation (1.2)

$$f(z_i) = \sum_{l=0}^{M} c_l P_l(z_i) + e_i \tag{1.2}$$

where z_i is the i[th] data point of z
$i = 0,1\cdots,N$
$l = 0,1\cdots,M$
$f(z_i)$ is the discrete data to be modeled
e_i is the error between $f(z_i)$ and the model at z_i

Equation (1.2) is the matrix equation (1.3)

$$\vec{b} = A\vec{x} + \vec{e} \tag{1.3}$$

Note the vectors in equation (1.3) are column vectors.

where $\vec{b} = f(z_i)$
A is an N×M matrix whose components are $a_{i,l} = P_l(z_i)$
$\vec{x}$ is a column vector of components c_l
$\vec{e}$ is a column vector of components e_i

Equation (1.4) clarifies equation (1.3)

$$\begin{bmatrix} f(z_0) \\ f(z_1) \\ \vdots \\ f(z_N) \end{bmatrix} = \begin{bmatrix} P_0(z_0) & P_1(z_0) & \cdots & P_M(z_0) \\ P_0(z_1) & P_1(z_1) & \cdots & P_M(z_1) \\ \vdots & \vdots & \cdots & \vdots \\ P_0(z_N) & P_1(z_N) & \cdots & P_M(z_N) \end{bmatrix} \cdot \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_M \end{bmatrix} + \begin{bmatrix} e_0 \\ e_1 \\ \vdots \\ e_M \end{bmatrix} \tag{1.4}$$

The columns of (1.4) are the basis functions at samples points z_i

Starting with equation (1.3) and solving for the error we get equation (1.5)

$$\vec{e} = \vec{b} - A\vec{x} \tag{1.5}$$

In the case where the measurements are not weighted equally, we get equation (1.6)

$$\vec{e}_1 = W\vec{e} = W\left(\vec{b} - A\vec{x}\right) \tag{1.6}$$

where W is a diagonal matrix of weights – $w_i = \dfrac{1}{\sigma_i}$ (1.7)

where σ_i is the standard deviation corresponding to each measurement - z_i

$$\begin{bmatrix} \frac{1}{\sigma_0} & 0 & 0 & 0 \\ 0 & \frac{1}{\sigma_1} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \frac{1}{\sigma_M} \end{bmatrix} \qquad (1.8)$$

Finding these weights apriori is not a trivial task. They can be estimated based on some physical attributes of the model or iterated. This paper will not go into further detail on this issue.

The estimator minimizes the squared error as shown in equation (1.9).

$$y = (W\vec{e})^T (W\vec{e}) = \vec{e}_1^T \vec{e}_1 = \vec{e}_1 \bullet \vec{e}_1 \qquad (1.9)$$

First we differentiate y with respect to $\vec{x}$ as shown in equation (1.10). Note that equation (1.10) differentiates a scalar with respect to a vector and uses the chain rule for differentiating a vector with respect to a vector. This chain rule is backwards from the normal calculus chain rule for differentiating a scalar with respect to a scalar[1].

$$\frac{dy}{d\vec{x}} = \frac{d\vec{e}}{d\vec{x}} \frac{d\vec{e}_1}{d\vec{e}} \frac{dy}{d\vec{e}_1} \qquad (1.10)$$

Equation (1.10) is by definition a column vector – see [1]. Equation (1.11) shows the differentiation of a dot product and is the standard definition of dot product differentiation. Also note that the derivative of a vector with respect to itself is the identity matrix – see definition of differentiating a vector with respect to a vector [1].

$$\frac{dy}{d\vec{e}_1} = \frac{d\vec{e}_1}{d\vec{e}_1} \bullet \vec{e}_1 + \vec{e}_1 \bullet \frac{d\vec{e}_1}{d\vec{e}_1} = I \bullet \vec{e}_1 + \vec{e}_1 \bullet I = 2\vec{e}_1 \qquad (1.11)$$

$$\frac{d\vec{e}_1}{d\vec{e}} = W^T \qquad (1.12)$$

Equation (1.12) is derived from the following:

If $\vec{s} = A\vec{x}$ or in index form, $s_i = a_{i,j} x_j$ (1.13)

We want to different $\vec{s}$ with respect to $\vec{x}$ - in index notation $\frac{d\vec{s}}{d\vec{x}} = \frac{ds_j}{dx_i}$, so we can change the indices of equation (1.13) so that there is an s_j and x_i as shown in equation (1.14)

$$s_j = a_{ji} x_i \tag{1.14}$$

Then differentiating (1.14) with respect to x_i gives

$$\frac{d\vec{s}}{d\vec{x}} = \frac{ds_j}{dx_i} = \frac{d\left(a_{ji} x_i\right)}{dx_i} = a_{ji} = A^T \tag{1.15}$$

Equation (1.15) can also be proved using the definition of differentiating a vector with respect to a vector. [1].

$$\frac{d\vec{e}}{d\vec{x}} = \frac{d}{d\vec{x}}\left(\vec{b} - A\vec{x}\right) = -A^T \tag{1.16}$$

Therefore,

$$\frac{dy}{d\vec{x}} = -A^T W^T 2\vec{e}_1 = -2A^T W^T W\vec{e} = -2A^T W^T W\left(\vec{b} - A\vec{x}\right) = 0 \tag{1.17}$$

Note that equation is a column vector as it should be.

Therefore

$$A^T W^T W\vec{b} = A^T W^T WA\vec{x} \tag{1.18}$$

$$A^T W^T W\vec{b} = A^T W^T WA\vec{x} \Rightarrow A^T C\vec{b} = A^T CA\vec{x} \Rightarrow \quad \hat{x} = \left(A^T CA\right)^{-1} A^T C\vec{b} \tag{1.19}$$

Equation (1.19) gives the solution for the least squares estimator

where $C = W^T W$

From equation (1.7)

$$C = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & & \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & \frac{1}{\sigma_M^2} \end{bmatrix} \tag{1.20}$$

The variance is given below:

$$\sigma^2 = E\left(e - E\left[e\right]^2\right) = E\left[e^2 - 2eE[e] + E[e]^2\right] = E[e^2] - 2E[e]E[e] + E^2[e] = E[e^2] - E^2[e] \tag{1.21}$$

Where E is the expected value and is defined below

$$E(x) = \int_{-\infty}^{\infty} x\, p(x)\, dx \tag{1.22}$$

$p(x) \equiv$ is a general probability distribution $\int_{-\infty}^{\infty} p(x)\, dx = 1$ (1.23)

If $E[e] = 0$ **Zero Mean of Measurement Errors** then

$$\sigma^2 = E[e^2] \tag{1.24}$$

This solution can be written in terms of an optimal estimation operator as follows

$\hat{x} = L_0 \vec{b}$, where $L_0 = \left(A^T C A\right)^{-1} A^T C$ (1.25)

L_0 is the optimal estimator for $\hat{x}$ given a measurement of $\vec{b}$

Symbolic Curve Fitting Examples

I.) This example uses a basis set of the first 4 Legendre Polynomials - P_0, P_1, P_2, P_3 **- and a diagonal** A **matrix**

We use the form for matrix A given by (1.26)

$$A = \begin{bmatrix} P_0 & 0 & 0 & 0 \\ 0 & P_1 & 0 & 0 \\ 0 & 0 & P_2 & 0 \\ 0 & 0 & 0 & P_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & x & 0 & 0 \\ 0 & 0 & \frac{3x^2-1}{2} & 0 \\ 0 & 0 & 0 & \frac{5x^3-3x}{2} \end{bmatrix} \tag{1.26}$$

We shall set the data vector $\vec{b}$ to be equation (1.27)

$$\vec{b} = \begin{bmatrix} 10P_0 \\ 7P_1 \\ 3P_2 \\ 1.5P_3 \end{bmatrix} = \begin{bmatrix} 10 \\ 7x \\ 3\frac{\left(3x^2-1\right)}{2} \\ 1.5\frac{5x^3-3x}{2} \end{bmatrix} \tag{1.27}$$

The weight vector will be the identity matrix $W = C = I$ and in this case $A^T = A$

$\left(A^T C A\right)^{-1}$ is given by equation (1.28)

$$\left(A^T C A\right)^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{x^2} & 0 & 0 \\ 0 & 0 & \frac{4}{\left(3x^2-1\right)^2} & 0 \\ 0 & 0 & 0 & \frac{4}{\left(5x^3-3x\right)^2} \end{bmatrix} \tag{1.28}$$

$\left(A^T C A\right)^{-1} A^T C$ is given by equation (1.29)

$$\left(A^T C A\right)^{-1} A^T C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{x} & 0 & 0 \\ 0 & 0 & \frac{2}{\left(3x^2-1\right)} & 0 \\ 0 & 0 & 0 & \frac{2}{\left(5x^3-3x\right)} \end{bmatrix} \tag{1.29}$$

Multiplying equation (1.29) by equation (1.27) we get equation (1.30)

$$\hat{x} = \left(A^T C A\right)^{-1} A^T C \vec{b} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{x} & 0 & 0 \\ 0 & 0 & \frac{2}{\left(3x^2-1\right)} & 0 \\ 0 & 0 & 0 & \frac{2}{\left(5x^3-3x\right)} \end{bmatrix} \bullet \begin{bmatrix} 10 \\ 7x \\ 3\frac{\left(3x^2-1\right)}{2} \\ 1.5\frac{5x^3-3x}{2} \end{bmatrix} = \begin{bmatrix} 10 \\ 7 \\ 3 \\ 1.5 \end{bmatrix} \tag{1.30}$$

which recovers the coefficients of the data of the Legendre Polynomials given in the data vector $\vec{b}$.

II.) **This example also uses a basis set of the first 4 Legendre Polynomials P_0, P_1, P_2, P_3 but uses a different configuration including off diagonal elements for the matrix A.**

The matrix A and the vector $\vec{b}$ are shown in equations (1.31) and (1.32).

$$A = \begin{bmatrix} & P_0 & P_1 & P_2 & P_3 \\ x^0 & 1 & 0 & -\frac{1}{2} & 0 \\ x^1 & 0 & x & 0 & -\frac{3x}{2} \\ x^2 & 0 & 0 & \frac{3x^2}{2} & 0 \\ x^3 & 0 & 0 & 0 & \frac{5x^3}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & -\frac{1}{2} & 0 \\ 0 & x & 0 & -\frac{3x}{2} \\ 0 & 0 & \frac{3x^2}{2} & 0 \\ 0 & 0 & 0 & \frac{5x^3}{2} \end{bmatrix} \tag{1.31}$$

$$\vec{b} = 5P_0 + 3P_2 + 4P_2 + 10P_3 = 25x^3 + 6x^2 - 12x + 3 = \begin{bmatrix} 3 \\ 12x \\ 6x^2 \\ 25x^3 \end{bmatrix} \tag{1.32}$$

Notice that the matrix A gives the Legendre polynomial basis set in terms of the polynomials $\left[x^0, x^1, x^2, x^3\right]$. The vector $\vec{b}$ also resolves the Legendre polynomial equation into the basis set $\left[x^0, x^1, x^2, x^3\right]$

$\left(A^T C A\right)^{-1}$ is given by equation (1.33)

$$\left(A^T C A\right)^{-1} = \begin{bmatrix} \frac{9x^4+1}{9x^4} & 0 & \frac{2}{9x^4} & 0 \\ 0 & \frac{25x^4+9}{25x^6} & 0 & \frac{6}{25x^6} \\ \frac{2}{9x^4} & 0 & \frac{4}{9x^4} & 0 \\ 0 & \frac{6}{25x^6} & 0 & \frac{4}{25x^6} \end{bmatrix} \tag{1.33}$$

$\left(A^T CA\right)^{-1} A^T C$ is given by equation (1.34)

$$\left(A^T CA\right)^{-1} A^T C = \begin{bmatrix} 1 & 0 & \frac{1}{3x^2} & 0 \\ 0 & \frac{1}{x} & 0 & \frac{3}{5x^3} \\ 0 & 0 & \frac{2}{3x^2} & 0 \\ 0 & 0 & 0 & \frac{2}{5x^3} \end{bmatrix} \tag{1.34}$$

Multiplying equation (1.34) and equation (1.32) yields equation (1.35)

$$\left(A^T CA\right)^{-1} A^T C\vec{b} = \begin{bmatrix} 1 & 0 & \frac{1}{3x^2} & 0 \\ 0 & \frac{1}{x} & 0 & \frac{3}{5x^3} \\ 0 & 0 & \frac{2}{3x^2} & 0 \\ 0 & 0 & 0 & \frac{2}{5x^3} \end{bmatrix} \bullet \begin{bmatrix} 3 \\ 12x \\ 6x^2 \\ 25x^3 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \\ 4 \\ 10 \end{bmatrix} \tag{1.35}$$

Equation (1.35) gives the correct Legendre polynomial coefficients – see equation (1.32)

References

[1]http://www.colorado.edu/engineering/CAS/courses.d/IFEM.d/IFEM.AppD.d/IFEM.AppD.pdf

[2] Strang, Gilbert, "Introduction to Applied Math", page 143, 1986