Constrained Linear Least Squares

CEE 201L. Uncertainty, Design, and Optimization
Department of Civil and Environmental Engineering
Duke University

Henri P. Gavin
Spring, 2015

The need to fit a curve to measured data arises in all branches of science, engineering, and economics. This hand-out addresses the ordinary least-squares method of fitting a curve to data in situations in which the curve-fit *must* satisfy certain criteria.

1 Unconstrained Ordinary Linear Least Squares

The fitting of an equation to data reduces to the estimation of unknown parameters in the equation from a set of data. If the parameters are multiplicative coefficients, e.g.,

$$\hat{y}(x; a_1, a_2, a_3, a_4, a_5, a_6) = a_1 + a_2 x + a_3 \sin(\pi x) + a_4 \cos(\pi x) + a_5 \exp(-x) + a_6 x^p \tag{1}$$

then the equation is linear in the coefficients, $\mathbf{a} = [a_1, \ldots, a_n]$. Given a set of m data points (x_1, y_1), ..., (x_m, y_m) (in which y_i is a measured response to the precise input x_i) the curve-fit equation can be evaluated m times, for each value of x_i, and in terms of the as-of-yet unknown coefficients: $(x_1, \hat{y}(x_1; \mathbf{a}))$, ..., $(x_m, \hat{y}(x_m; \mathbf{a}))$. For curve-fit equations which are linear in the coefficients, the m equations can be written in matrix form, for example,

$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_m \end{bmatrix} = \begin{bmatrix} 1 & x_1 & \sin(\pi x_1) & \cos(\pi x_1) & \exp(-x_1) & x_1^p \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & \sin(\pi x_m) & \cos(\pi x_m) & \exp(-x_m) & x_m^p \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_6 \end{bmatrix}$$

or, in matrix-notation short hand, $\hat{\mathbf{y}} = \mathbf{X}\mathbf{a}$. Note that the matrix $\mathbf{X}$ depends *only* on the precisely-known values of the independent variables, x_i. The columns of $\mathbf{X}$ can be linear in x or non-linear in x. This matrix can not contain values of the measured dependent variable y_i, nor can it contain any coefficient, a_i. In most curve-fitting problems, there are more data points than curve-fit coefficients ($m > n$) so there are more equations than unknown; the matrix equation $\hat{\mathbf{y}} = \mathbf{X}\mathbf{a}$ represents an *over-determined* system of linear equations.

Using the least-squares approach to estimating the curve-fit coefficients, $\mathbf{a}$, the objective function to be minimized is the sum of the squares of the differences between the curve-fit equation, $\hat{y}_i$ and the associated measured data y_i,

$$J(\mathbf{a}) = \sum_{i=1}^{m} \left[\ \hat{y}(x_i; \mathbf{a}) - y_i \ \right]^2 \tag{2}$$

This sum-of-squares-of-errors ("SSE") criterion can be written with matrix notation

$$\begin{aligned} J(\mathbf{a}) &= (\hat{\mathbf{y}} - \mathbf{y})^{\mathsf{T}} (\hat{\mathbf{y}} - \mathbf{y}) \\ &= (\mathbf{Xa} - \mathbf{y})^{\mathsf{T}} (\mathbf{Xa} - \mathbf{y}) \\ &= (\mathbf{a}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}} - \mathbf{y}^{\mathsf{T}}) (\mathbf{Xa} - \mathbf{y}) \\ &= \mathbf{a}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{Xa} - \mathbf{a}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y} - \mathbf{y}^{\mathsf{T}}\mathbf{Xa} + \mathbf{y}^{\mathsf{T}}\mathbf{y} \end{aligned} \tag{3}$$

Note that the matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is the Hessian matrix of this quadratic objective function. As long as more than n columns of $\mathbf{X}$ are linearly independent, this Hessian matrix is guaranteed to be positive-definite ($\mathbf{a}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{Xa} > 0 \; \forall \; \mathbf{a} \neq \mathbf{0}$) and, therefore, invertible. The necessary criterion for minimizing J with respect to the set of curve-fit coefficients, $\mathbf{a}$, is that the gradient of J with respect to the unknown coefficients, $\mathbf{a}$, must be zero.

$$\frac{\partial J}{\partial \mathbf{a}} = \mathbf{0} \iff \mathbf{0} = \mathbf{a}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X} + (\mathbf{X}^{\mathsf{T}}\mathbf{Xa})^{\mathsf{T}} - (\mathbf{y}^{\mathsf{T}}\mathbf{X}) - (\mathbf{X}^{\mathsf{T}}\mathbf{y})^{\mathsf{T}} \iff \mathbf{X}^{\mathsf{T}}\mathbf{Xa} = \mathbf{X}^{\mathsf{T}}\mathbf{y}$$

and the ordinary least-squares estimates for the curve-fit coefficients can be computed from

$$\mathbf{a}^* = [\mathbf{X}^{\mathsf{T}}\mathbf{X}]^{-1} \; \mathbf{X}^{\mathsf{T}}\mathbf{y}. \tag{4}$$

2 Constrained Ordinary Linear Least Squares

Now, suppose that in addition to minimizing the sum-of-squares-of-errors, the curve-fit must also satisfy other criteria. For example, suppose that the curve-fit must pass through a particular point (x_c, y_c), or that the slope of the curve at a particular location, x_s, must be exactly a given value, y'_s. Satisfying such constraints is a natural application of the method of Lagrange multipliers. Using the example started above, these two criteria can be written,

$$y_c = \hat{y}(x_c; \mathbf{a}) = a_1 + a_2 x_c + a_3 \sin(\pi x_c) + a_4 \cos(\pi x_c) + a_5 \exp(-x_c) + a_6 x_c^p$$

and

$$y'_s = \hat{y}'(x_c; \mathbf{a}) = 0a_1 + a_2 + a_3 \pi \cos(\pi x_s) - a_4 \pi \cos(\pi x_s) - a_5 \exp(-x_s) + a_6 p x_c^{p-1},$$

or, in matrix form,

$$\left[\begin{array}{c} y_c \\ y'_s \end{array} \right] = \left[\begin{array}{cccccc} 1 & x_c & \sin(\pi x_c) & \cos(\pi x_c) & \exp(-x_c) & x_c^p \\ 0 & 1 & \pi \cos(\pi x_s) & -\pi \sin(\pi x_s) & -\exp(-x_s) & p x_s^{p-1} \end{array} \right] \left[\begin{array}{c} a_1 \\ \vdots \\ a_6 \end{array} \right]$$

or, in short hand, $\mathbf{b} = \mathbf{Aa}$. So now the problem is to minimize $J(\mathbf{a})$ (equation (3)) such that $\mathbf{Aa} = \mathbf{b}$. This is a linearly-constrained quadratic minimization ... an *ideal* problem for Lagrange multipliers. The augmented objective function (the *Lagrangian*) is then,

$$J_A(\mathbf{a}, \boldsymbol{\lambda}) = \mathbf{a}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{Xa} - \mathbf{a}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y} - \mathbf{y}^{\mathsf{T}}\mathbf{Xa} + \mathbf{y}^{\mathsf{T}}\mathbf{y} + \boldsymbol{\lambda}^{\mathsf{T}}(\mathbf{Aa} - \mathbf{b}) \tag{5}$$

Minimizing J_A with respect to $\mathbf{a}$ and maximizing J_A with respect to $\boldsymbol{\lambda}$ results in a system of linear equations for the optimum coefficients $\mathbf{a}^*$ and Lagrange multipliers $\boldsymbol{\lambda}^*$.

$$\begin{bmatrix} 2\mathbf{X}^\mathsf{T}\mathbf{X} & \mathbf{A}^\mathsf{T} \\ \mathbf{A} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{a}^* \\ \boldsymbol{\lambda}^* \end{bmatrix} = \begin{bmatrix} 2\mathbf{X}^\mathsf{T}\mathbf{y} \\ \mathbf{b} \end{bmatrix} \tag{6}$$

If the curve-fit problem has n coefficients and c constraint equations, then the matrix is square and of size $(n + c) \times (n + c)$.

3 Example

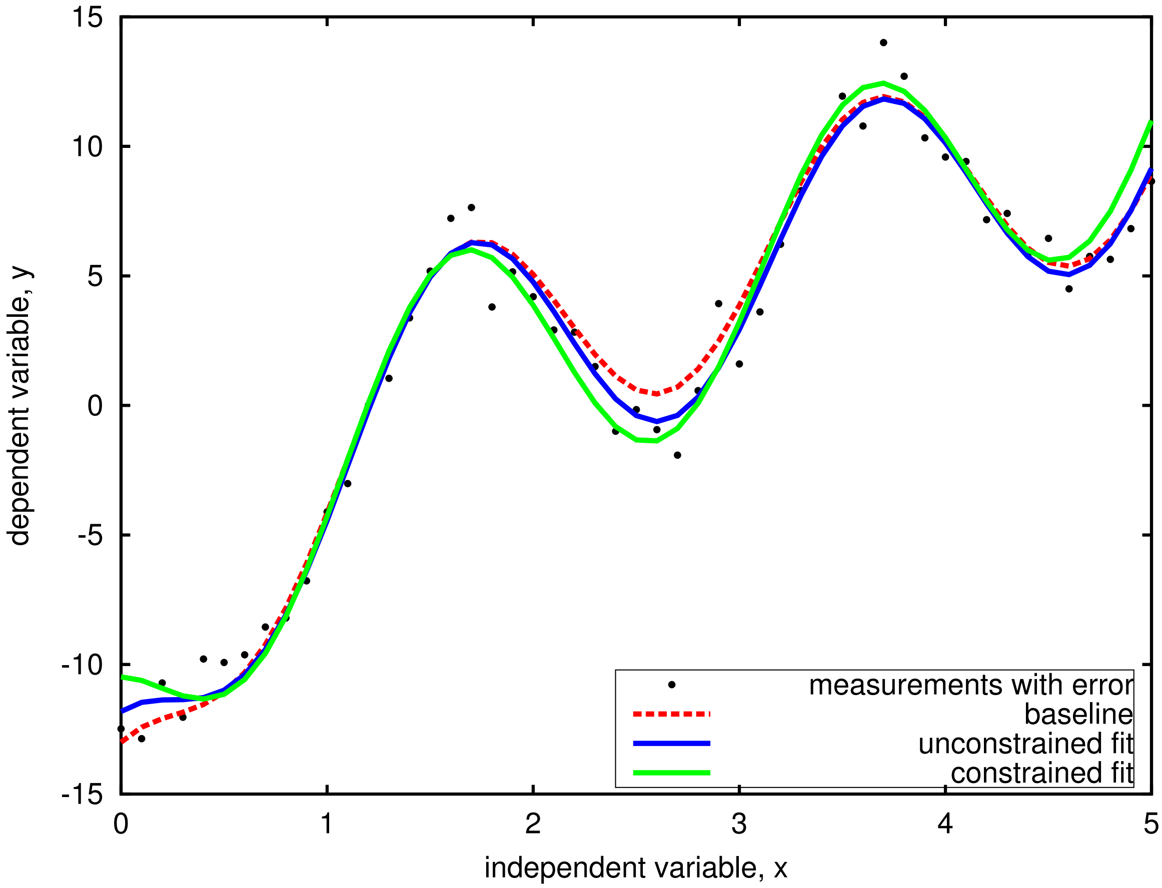
As an example, consider fitting equation (1) to a set of data points. An assumption implicit to the least-squares error criterion (equation (2)) is that the independent variables x_i have no error, and that all the error is in the y_i values. To illustrate the method, the curve-fit equation is fit to m noisy data points. The data points are generated numerically by evaluating the fit equation at m values of x_i with a set of "true" pre-selected coefficients, and subsequently adding random numbers to the data. A set of coefficients can then be estimated from the noisy data and compared to the "true" coefficient values.

In this example, $x_i \in [0.0, 0.1, 0.2, \cdots, 5]$; the "true" (pre-selected) coefficient values are given in the table below. The exponent value is given as $p = 2.5$. Note that including this exponent as a curve-fit coefficient would make the problem *non-linear* in the coefficients, and it could not be solved using linear least squares methods. The measurement errors are simulated as being normally-distributed with a mean of zero and a standard deviation of 1.

For the constrained fit, $\hat{y}(5)$ is constrained to have a value of 11, and $\hat{y}'(0)$ is constrained to have a value of 0.

The unconstrained curve-fit is found from equation (4) and the constrained curve-fit is found from equation (6). The resulting parameter values are shown in the table below, and the related data and curves are shown in the figure. Note that the estimated parameters depend on both the underlying trend (the pre-selected coefficient values) *and* the random errors that are added-in.

coefficient	pre-selected "true"	unconstrained estimate	constrained estimate
a_1	5.00	5.46	3.06
a_2	0.10	-0.62	0.02
a_3	-4.00	-4.32	-4.91
a_4	2.00	2.26	1.87
a_5	-20.00	-19.53	-15.40
a_6	0.10	0.16	0.18



The coefficient estimates $(a_1^*, \cdots, a_6^*)$ are computed from 51 data points (black points). The data are generated by adding normally-distributed random values to a baseline curve (dashed red). The unconstrained fit is plotted with a blue line and the constrained fit is plotted with a green line. Note that the constrained fit satisfies the criteria $\hat{y}'(0) = 0$ and $\hat{y}(5) = 11$. Otherwise, both fits minimize the sum-of-the-squares-of-the-errors ("SSE") criterion.

For smaller levels of measurement error, and for data sets with more data points, the estimated unconstrained fit coefficients would match the "true" values more precisely. However, the estimated constrained fit coefficients should not match the "true" values, no matter how small the measurement error or how many data points, since the fit function with the "true" values do not satisfy the constraints.