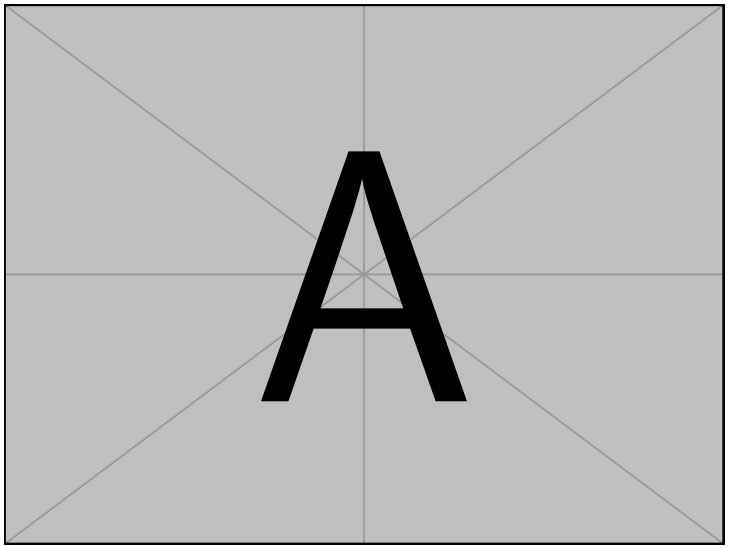
CEE384—Numerical Methods

Regression

Least Square Regression



Arizona State University

Brian Chevalier

Updated: November 27, 2019

Contents

1 Linear Regression **1**

2 General Polynomial Regression **2**

References **3**

1 Linear Regression

A general linear equation takes the form:

$$y = a_0 + a_1 x \tag{1}$$

where y is the function output, a_0 is the constant term, a_1 is the slope, and x is the function input. This equation can be used to interpolate between two points, or create a series of linear spline interpolation equations. However, here we will be looking not at precisely crossing all points in a data set, but instead, minimizing the error associated with creating a linear model for a set of data.

The data are a series of x and y coordinates in the form shown in the following equations:

$$\mathbf{x} = \begin{bmatrix} x_0 & x_1 & \dots & x_{n-1} \end{bmatrix} \tag{2}$$

$$\mathbf{y} = \begin{bmatrix} y_0 & y_1 & \dots & y_{n-1} \end{bmatrix} \tag{3}$$

where n is the number of elements in the data set.

The linear regression equation is:

$$y = a_0 + a_1 x + e \tag{4}$$

where e is the error (sometimes called the residual), or the difference between the equation and the actual data set. Now, we want to *minimize* the sum of the squares of the error.

$$S_r = \sum_{k=0}^{k<n} (e_k)^2 = \sum_{k=0}^{k<n} (y_k - a_0 - a_1 x_k)^2 \tag{5}$$

where the subscript k denotes the kth element of the x and y data set. Taking the derivative with respect to each coefficient term yields:

$$\frac{\partial S_r}{\partial a_0} = -2 \sum_{k=0}^{k<n} (y_k - a_0 - a_1 x_k) = 0 \tag{6}$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum_{k=0}^{k<n} x_k (y_k - a_0 - a_1 x_k) = 0 \tag{7}$$

Distributing the sum and rearranging:

$$n a_0 + a_1 \sum x_k = \sum y_k \tag{8}$$

$$a_0 \sum x_k + a_1 \sum x_k^2 = \sum x_k y_k \tag{9}$$

Note that the bounds on the summations have been dropped for the sake of brevity and is short for $\sum_{k=0}^{k<n}$.

Putting the system of equations in matrix form:

$$\begin{bmatrix} n & \sum x_k \\ \sum x_k & \sum x_k^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \end{Bmatrix} = \begin{Bmatrix} \sum y_k \\ \sum x_k y_k \end{Bmatrix} \tag{10}$$

Solving the system of equations yields:

$$a_1 = \frac{n \sum x_k y_k - \sum x_k \sum y_k}{n \sum x_k^2 - \left(\sum x_k \right)^2} \tag{11}$$

$$a_0 = \bar{y} - a_1\bar{x} \tag{12}$$

where $\bar{y} = \frac{\sum y_k}{n}$ and $\bar{x} = \frac{\sum x_k}{n}$.

2 General Polynomial Regression

To establish the general n-order polynomial equations, we will derive the quadratic equations and find the pattern between these and the linear equations. The quadratic regression formulas can be derived through the same method as their linear counterparts.

$$y = a_0 + a_1 x + a_2 x^2 + e \tag{13}$$

$$S_r = \sum_{k=0}^{k<n} \left(y_k - a_0 - a_1 x_k - a_2 x_k^2\right)^2 \tag{14}$$

Taking the derivative with respect to each coefficient:

$$\frac{\partial S_r}{\partial a_0} = -2 \sum \left(y_k - a_0 - a_1 x_k - a_2 x_k^2\right) \tag{15}$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum x_k \left(y_k - a_0 - a_1 x_k - a_2 x_k^2\right) \tag{16}$$

$$\frac{\partial S_r}{\partial a_2} = -2 x_k^2 \sum x_k \left(y_k - a_0 - a_1 x_k - a_2 x_k^2\right) \tag{17}$$

Expanding the equations:

$$a_0 n + a_1 \sum x_k + a_2 \sum x_k^2 = \sum y_k \tag{18}$$

$$a_0 \sum x_k + a_1 \sum x_k^2 + a_2 \sum x_k^3 = \sum x_k y_k \tag{19}$$

$$a_0 \sum x_k^2 + a_1 \sum x_k^3 + a_2 \sum x_k^4 = \sum x_k^2 y_k \tag{20}$$

$$\underbrace{\begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} n & \sum x_k & \sum x_k^2 \\ \sum x_k & \sum x_k^2 & \sum x_k^3 \\ \sum x_k^2 & \sum x_k^3 & \sum x_k^4 \end{bmatrix} \end{array}}_{\mathbf{A}} \underbrace{\begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix}}_{\mathbf{x}} = \underbrace{\begin{Bmatrix} \sum y_k \\ \sum x_k y_k \\ \sum x_k^2 y_k \end{Bmatrix}}_{\mathbf{b}} \tag{21}$$

Matching the pattern. There is a very clear pattern for the elements other than $A_{0,0}$ and b_0. However, replacing element $A_{0,0}$ with $\sum x_k^0$, and element b_0 with $\sum x_k^0 y_k$, creates a much clearer pattern to follow. The system of equations becomes:

$$\underbrace{\begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array}\begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} \sum x_k^0 & \sum x_k^1 & \sum x_k^2 \\ \sum x_k^1 & \sum x_k^2 & \sum x_k^3 \\ \sum x_k^2 & \sum x_k^3 & \sum x_k^4 \end{bmatrix} \end{array}}_{\mathbf{A}} \underbrace{\begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix}}_{\mathbf{x}} = \underbrace{\begin{Bmatrix} \sum x_k^0 y_k \\ \sum x_k^1 y_k \\ \sum x_k^2 y_k \end{Bmatrix}}_{\mathbf{b}} \tag{22}$$

Building the equations in a matrix equation can be summarized with the following two equations:

$$A_{i,j} = \sum_{k=0}^{k<n} x_k^{i+j} \quad (0 \leq i < n) \quad (0 \leq j < n) \tag{23}$$

where i is the row index, and j is the column index.

$$b_i = \sum_{k=0}^{k<n} x_k^i y_k \quad (0 \leq i < n) \tag{24}$$

These equations can now be programmed, and solved using linear algebra techniques (i.e. gauss elimination).

References

[1] Chapra, Steven C. and Canale, Raymond P., *Numerical Methods for Engineers*, 7th ed. McGraw Hill Education, 2015.

[2] A. K. Kaw, E. E. Kalu, and D. Nguyen. Numerical methods with applications. [Online]. Available: http://nm.mathforcollege.com/topics/textbook_index.html