Computational Maths - Chapter 6

Varjak Wolfe

November 12, 2021

This follows Chapter 6 of the textbook.

-Curve Fitting

-Interpolation: Using the known points to estimate expected values between them.

-Linear Least Squares Regression

Curve Fitting a Linear Equation

$y = a_1x+a_0$ is used to best fit given data points. This is done by determining the constants a_1 and $a0$ that give the smallest error when the data points are substituted in.

The process of obtaining the constants that give the best fit requires us to have a definition of best fit and a mathematical procedure for deriving the value of the constants.

The fit between given data points and an approximating linear function is determiend by first calculating the error (also called the residual) which is the difference between a data point and the value of the approximating function, at each point. So the residuals are used for calculating a total error for all the points.

The Residual r_i at a point (x_i, y_i) is the difference between the value y_i of the data point and the value of the function $f(x_i)$ used to approximate the data points:

$$r_i = y_i - f(x_i)$$

To measure how well the approximating function fits the given data can be obtained by calculating a total error E in terms of the residuals.

Add all the absolute values of the residuals of all the points, because positive and negative residuals can cancel each other out and give an inancurate measure of fit, and also we need a formula that can give us a unique linear function that has the best fit (smallest error) is obtained by:

$$E = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} [y_i - (a_1x_i + a_0)]^2$$

Linear Least Squares Regression

This is a procedure in which the coefficients $a_0 and a_1$ of a linear function $y = a_1x + a_0$ are determined such that the function has the best fit to a given

set of data points. The best fit is defined as the smallest possible total error that is calculated by adding the squares of the residuals according to

$$E = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} [y_i - (a_1 x_i + a_0)]^2$$

Calculate the summations first:
$S_x = \sum_{i=1}^{n} X_i$
$S_y = \sum_{i=1}^{n} Y_i$
$S_{xx} = \sum_{i=1}^{n} X_i^2$
$S_{xy} = \sum_{i=1}^{n} X_i Y_i$
and then substitute them in to the equations for a_0 and a_1

$$a_1 = \frac{nS_{xy} - S_x S_y}{nS_{xx} - (S_x)^2}$$

$$a_0 = \frac{S_{xx} S_y - S_{xy} S_x}{nS_{xx} - (S_x)^2}$$

Example 6-1

T(°C)	0	10	20	30	40	50	60	70	80	90	100
p (atm.)	0.94	0..96	1.0	1.05	1.07	1.09	1.14	1.17	1.21	1.24	1.28

Use linear least-squares regression to determine a linear function in the form p = a1 T + a0 that best fits the data points. Use 4 data points.
$S_x = \sum_{i=1}^{4} X_i = 0 + 30 + 70 + 100 = 200$
$S_y = \sum_{i=1}^{4} Y_i = 0.94 + 1.05 + 1.17 + 1.28 = 4.44$
$S_{xx} = \sum_{i=1}^{4} X_i^2 = 0^2 + 30^2 + 70^2 + 100^2 = 15800$
$S_{xy} = \sum_{i=1}^{4} X_i Y_i = (0 * 0.94) + (30 * 1.05 + (70 * 1.17) + (100 * 1.28) = 241.4$
Substituting the Ss

$$a_1 = \frac{nS_{xy} - S_x S_y}{nS_{xx} - (S_x)^2} = \frac{(4 * 241.4) - (200 * 4.44)}{(4 * 15800) - (200)^2} = 0.003345$$

$$a_0 = \frac{S_{xx} S_y - S_{xy} S_x}{nS_{xx} - (S_x)^2} = \frac{(15800 * 4.44) - (241.2 * 200)}{(4 * 15800 - (200)^2} = 0.9428$$

The equation of best fit is $y = a_1 x + a_0 = p = 0.003345T + 0.9428$

Curve Fitting with Nonlinear Equation by Writing it in Linear Form

Sometimes it can be clear that curve fitting the data points with a nonlinear function guces a better fit than curve fitting with a linear function. We will look at curve fitting with nonlinear functions that can be written in a form for which

the Linear Least Squares Regression method can still be used for determining the coefficients that give the best fit.

Examples such as:

$y = bx^m$ power

$y = be^{mx}$ or $y = b10^{mx}$ exponential

$y = \frac{1}{mx+b}$ reciprocal

Writing a Nonlinear Equation in Linear Form

$y - bx^m$ can be put into linear form by taking the natural log of both sides

$$ln(y) = ln(bx^m) = mln(x) + ln(b)$$

This equation is linear for ln(y) in terms of $ln(x)$. The equation is of the form $y = a_1x + a_0$ where $y = ln(y)$, $a_1 = m$, $x = ln(x)$ and $a_0 = ln(b)$

So we can do the Linear Least Squares Regression by substituting $ln(y_i)$ for y_i and $ln(x_i)$ for x_i.

Once a_1 and a_0 are known. constants b and m in the exponential equation are calculated by

$$m = a_1, b = e^{a_0}$$

Curve Fitting with Quadratic and Higher Order Polynomials

Polynomials are functions of the form:

$$f(x) = a_nx^n + a_{n-1}x^{n-1} + ... + a_1x + a_0$$

n is the degree/order of the polynomial. A plot of the polynomial is a curve. First order polynomial is linear and its plot is a straight line. Higher order polynomials are nonlinear functions with plots that curve. A quadratic (second order) polynomial is a curve that is either concave up or down (parabola). A third order has an inflection point such that the curve can be concave up or down in one region and concave down or up in another.

A given set of n data points can be curve-fit with polynomials of different order up to an order of (n-1). The coefficients can be determined such that the polynomial best fits the data by minimizing the error in a least squares sense.

For any number of data points n, it is possible to derive a polynomial (order of n-1) that apsses exactly through all the points. However there is large deviation between the points. So even though a higher order polynomial gives the exact values at all the data points, it can not be relied upon for accurate interpolaton or extrapolation.

Polynomial Regression

This is a procedure for determining the coefficients of a polynomial of a second degree, or higher, such that the polynomial best fits a given set of data points. As in linear regression, the derivation of the equations is base don minimizing total error.

If the polynomial, of order m, that is used for the curve fitting is:

$$f(x) = a_mx^m + a_{m-1}x^{m-1} + ... + a_1x + a_0$$

then, for a given set of n data points (x_i, y_i) (m is smaller than n-1), the total error is given by:

$$E = \sum_{i=1}^{n} [y_i - (a_m x^m + a_{m-1} x^{m-1} + ... + a_1 x + a_0)]^2$$

Since all the values x_i and y_i of the data points are known, E is a nonlinear function of the m+1 variables (the coefficients a_0 through a_m). E has a minimum at the values of a_0 through a_m where the partial deriivatives of E with respect to each of the variables is equal to zero. Taking the partial derivatives of E and setting them to zero gives a set of m+1 linear equations for the coefficients.

Taking an example for the derivation of case m = 2, a quadratic polynomial. In that case, our E function is:

$$E = \sum_{i=1}^{n} [y_i - (a_2 x_i^2 + a_1 x_i + a_0)]^2$$

Taking the partial derivatives with respect to a_0, a_1, a_2 and setting them equal to zero gives:

$$\frac{\partial E}{\partial a_0} = -2 \sum_{i=1}^{n} (y_i - a_2 x_i^2 + a_1 x_i + a_0) = 0$$

$$\frac{\partial E}{\partial a_1} = -2 \sum_{i=1}^{n} (y_i - a_2 x_i^2 + a_1 x_i + a_0) x_i = 0$$

$$\frac{\partial E}{\partial a_2} = -2 \sum_{i=1}^{n} (y_i - a_2 x_i^2 + a_1 x_i + a_0) x_i^2 = 0$$

This is now a system of three linear equations for the unknowns a_0, a_1, a_2 which can be rewritten:

$$n a_0 + (\sum_{i=1}^{n} x_i) a_1 + (\sum_{i=1}^{n} x_i^2) a_2 = \sum_{i=1}^{n} y_i$$

$$(\sum_{i=1}^{n} x_i) a_0 + (\sum_{i=1}^{n} x_i^2) a_1 + (\sum_{i=1}^{n} x_i^3) a_2 = \sum_{i=1}^{n} x_i y_i$$

$$(\sum_{i=1}^{n} x_i^2) a_0 + (\sum_{i=1}^{n} x_i^3) a_1 + (\sum_{i=1}^{n} x_i^4) a_2 = \sum_{i=1}^{n} x_i^2 y_i$$

The solution of this system of equations gives the values of the coefficients a_0, a_1, a_2 of the polynomial $y = a_2x_i^2 + a_1x_i + a_0$ that best fits the n data points (x_i, y_i)

Interpolation using a Single Polynomial

Interpolation is a procedure in which a formula is used to represent a given set of data points, such that the formula gives the exact value at all the data points and an estimated value between the points.

As said previously, for any number of points n there is a polynomial of order n-1 that passes through all of the points.

For two points, the polynomial is of first order (a straight line connecting all the points). For three points, the polynomial is of second order (a parabola that connects the points) and so on.

Once the polynomial is determined, it can be used for estimating the y values between the known points simply by substituting for the x coordinate in the polynomial.

To fix the problem we mentione dearlier where high order polynomials deviate a lot between the points and thus cant be used reliably for interpolation between the points; we use piecewise interpolation in which different lower-order polynomials are used for interpolation between points.

For a given set of n points, only one unique polynomial of order n-1 passes exactly through all of the points. However, the polynomial can be written in different forms. We will now see how to derive three forms of polynomials (standard, lagrange and newton's). The different forms are suitable for different circumstances.

The standard form we have seen already:

$$f(x) = a_nx^n + a_{n-1}x^{n-1} + ... + a_1x + a_0$$

The coefficents in this form are determined by solving a system of m+1 linear equations. The equations are obtained by writing the polynomial explicitly for each point (substituting each point into the polynomial).

Foe example, the n = 5 points in the fourth degree m = 4 polynomial are: (1,2), (4,6), (7,4), (10,8) and (13,10).

Writing a polynomial in standard form for each point gives the following system of equations for the unknowns $a_0 to a_4$:

$$f(x) = a_4(1)^4 + a_3(1)^3 + a_2(1)^2 + a_1(1) + a_0 = 2$$

$$f(x) = a_4(4)^4 + a_3(4)^3 + a_2(4)^2 + a_1(4) + a_0 = 6$$

$$f(x) = a_4(7)^4 + a_3(7)^3 + a_2(7)^2 + a_1(7) + a_0 = 4$$

$$f(x) = a_4(10)^4 + a_3(10)^3 + a_2(10)^2 + a_1(10) + a_0 = 8$$

$$f(x) = a_4(13)^4 + a_3(13)^3 + a_2(13)^2 + a_1(13) + a_0 = 10$$

The solution of this system gives the values of the coefficients. Done by MATLAB:

```
>> a = [1 1 1 1 1; 4^4 4^3 4^2 4 1; 7^4 7^3 7^2 7 1; 10^4 10^3
10^2 10 1; 13^4 13^3 13^2 13 1]
a =
        1         1         1         1         1
      256        64        16         4         1
     2401       343        49         7         1
    10000      1000       100        10         1
    28561      2197       169        13         1
>> b = [2; 6; 4; 8; 10]
>> A = a\b
A =
   -0.0103
    0.3004
   -2.8580
   10.1893
   -5.6214
```

The polynomial that corresponds to these coefficients is:
$y = -0.0103x^4 + 0.3x^3 - 2.86x^2 + 10.19x - 5.62$
(see Fig. 6-11).

We will now look at forms of writing the polynomials so that they are easier to use

Lagrange Interpolating Polynomials

For two points $(x_1, y_1 0 and (x_2, y_2)$, the first-order Lagrange polynomial that apsses through the points has the form:

$$f(x) = y = a_1(x - x_2) + a_2(x - x_1)$$

Substituting the two points into this gives:

$$y_1 = a_1(x_1 - x_2) + a_2(x_1 - x_1)$$

or $a_1 = \frac{y_1}{(x_1 - x_2)}$

and

$$y_2 = a_1(x_2 - x_2) + a_2(x_2 - x_1)$$

or $a_2 = \frac{y_2}{(x_2 - x_1)}$

Subsitituting the coefficients a_1 and a_2 back into the original gives:

$$f(x) = y = a_1(x - x_2) + a_2(x - x_1)$$

$$f(x) = \frac{(x - x_2)}{(x_1 - x_2)})y_1 + \frac{(x - x_1)}{(x_2 - x_1)})y_2$$

This is a linear function of x. If $x = x_1$ is substituted into it, the value of the polynomial is y_1 and if $x = x_2$ is subbed in, the value is y_2. Subbing a value of x between the points gives an interpolated value of y.

This equations can also be rewritten in the standard form $f(x) = a_1x + a_0$:

$$f(x) = \frac{(y_2 - y_1)}{(x_2 - x_1)})x + \frac{x_2y_1 - x_1y_2}{(x_2 - x_1)}$$

For three points $(x_1, y_1), (x_2, y_2), (x_3, y_3)$, the second-order Lagrange polynomial that passes through the points has the form:

$$f(x) = \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)}y_1 + \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)}y_2 + \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)}y_3$$

This equation is a quadratic function of x. When the coordinate x_1, x_2 or x_3 of one of the three given points is substitiuted into the above equation, the value of the polynomial is equal to y_1, y_2 or y_3, respectively. This is because the coefficient in front of the corresponding y_i is equal to 1 and the coefficient of the other two terms is equal to zero.

Newton Interpolating Polynomials (Newtonian Divided Difference)

For three given points, the second-order Newton's polynomial has the form:

$$f(x) = a_1 + a_2(x - x_1) + a_3(x - x_1)(x - x_2)$$

It is an equation of a parabola that passes through the three points. The coefficients a_1, a_2, a_3 can be determined by subbing the three points into this equation.

Subbing $x = x_1$ and $f(x_1) = y_1$ gives $a_1 = y_1$. Subbing the second point $x = x_2$ and $f(x_2) = y_2$ (and $a_1 = y_1$ gives:

$$y_2 = y_1 + a_2(x_2 - x_1)$$

or

$$a_2 = \frac{y_2 - y_1}{x_2 - x_1}$$

Subbing the third point $x = x_3$ and $f(x_3) = y_3$ (as well as $a_1 = y_1$ and $a_2 = \frac{y_2 - y_1}{x_2 - x_1}$ gives:

$$y_3 = y_1 + \frac{y_2 - y_1}{x_2 - x_1}(x_3 - x_1) + a_3(x_3 - x_1)(x_3 - x_2)$$

This equation can be solved for a_3 and rearranged to give:

$$a_3 = \frac{\frac{y_3 - y_2}{x_3 - x_2} - \frac{y_2 - y_1}{x_2 - x_1}}{(x_3 - x_1)}$$