Model Fitting *

CURM Background Material, Fall 2014
Dr. Doreen De Leon

1 Introduction

Given a set of data points, we often want to fit a selected model or type to the data (e.g., we suspect an exponential behavior), and then choose the most appropriate model by determining which model fits best.

Before discussing criteria on which to base curve-fitting decisions, we first will look at sources and types of errors.

Formulation errrors. These errors result from the assumption that certain variables are negligible or from simplifications in describing interrelationships among the variables in the various submodels constructed as part of the modeling process.

Truncation errors. These errors are attributable to numerical methods used to solve a mathematical problem. An example is the use of a Taylor polynomial to approximate a function.

Round-off errors. These errors are caused by using finite-digit machines for computations. Each number can be represented in a computer with a number having only finite precision. Thus, each arithmetic operation that is performed, each with its own round-off, causes an accumulated round-off error. Such an error can significantly alter the results of a model's solution.

Measurement errors. These errors are caused by imprecision in the data collection. This may result from human error in recording data or from limitations in the accuracy of measuring equipment.

2 Analytic Methods of Model Fitting

There are actually several methods for fitting curves to a set of data points, based on different sets of criterion. We will briefly examine some of these, before focusing on the most common.

*The majority of the material in this handout is taken directly from *A First Course in Mathematical Modeling* by Frank Giordano, et al.

Note that this may also be found in Section 4.1 of the textbook.

Least-Squares Criterion

Currently, the most frequently used curve-fitting criterion is the least-squares criterion. In this case, given some function type $y = f(x)$ and a set of m data points $(x_i, y_i), i = 1, 2, \ldots, m$, we seek to minimize the sum

$$\sum_{i=1}^{m} |y_i - f(x_i)|^2. \tag{1}$$

In other words, we are minimizing the standard deviation of the function $f(x)$ from the data set.

One of the advantages of the least-squares method (also known as regression) is that the resulting optimization problem can be solved only using the calculus of several variables.

Minimizing the Sum of the Absolute Deviations

In this case, given some function type $y = f(x)$ and a set of m data points $(x_i, y_i), i = 1, 2, \ldots, m$, we seek to minimize

$$\sum_{i=1}^{m} |y_i - f(x_i)|.$$

Again, in general, this presents some problems because, in order to solve, we must differentiate the above sum with respect to the parameters in $f(x_i)$ to find the critical points. However, because of the absolute values, the derivatives are not continuous.

Chebyshev Approximation Criterion

Given some function type $y = f(x)$ and a set of m data points $(x_i, y_i), i = 1, 2, \ldots, m$, minimize the largest absolute deviation $|y_i - f(x_i)|$ over the entire set.

The difficulty with the Chebyshev criterion is that it is often difficult to apply in practice, because application of the criterion results in an optimization problem that may require advanced mathematical procedures or numerical algorithms to solve.

Relating the Criteria

The geometric interpretations of the three curve-fitting criteria help provide a qualitative description comparing the criteria. Minimizing the sum of the absolute deviations tends

to treat each data point with equal weight and to average the deviations. The Chevyshev criterion gives more weight to a single point potentially having a large deviation. The least-squares criterion, in terms of weighting individual points, is somewhat in between both of these criteria.

The following also may be found in Section 4.3 of the tetbook. Now, we will compare the Chebyshev and least-squares approaches. Suppose the Chebyshev criterion is applied and the resulting optimization problem solved to give the function $f_1(x)$. The absolute deviations resulting from the fit are defined as follows:

$$|y_i - f_1(x_i)| = c_i, i = 1, 2, \ldots, m.$$

Define c_{max} as the largest of the absolute deviations c_i. Since the parameters of $f_1(x)$ are determined so as to minimize c_{max}, it is the minimal largest absolute deviation obtainable.

Suppose, instead, that the least-squares criterion is applied and the resulting optimization problem solved to yield the function $f_2(x)$. The absolute deviations resulting from this fit are given by

$$|y_i - f_2(x_i)| = d_i, i = 1, 2, \ldots, m.$$

Define d_{max} as the largest of the absolute deviations d_i. Clearly, $d_{max} \geq c_{max}$. Since the sum of the squares of d_i is the smallest such sum obtainable, we know that

$$d_1^2 + d_2^2 + \cdots + d_m^2 \leq c_1^2 + c_2^2 + \cdots + c_m^2.$$

However, $c_i \leq c_{max}$ for $i = 1, 2, \ldots, m$. Therefore,

$$d_1^2 + d_2^2 + \cdots + d_m^2 \leq m c_{max}^2,$$

or

$$D = \sqrt{\frac{d_1^2 + d_2^2 + \cdots + d_m^2}{m}} \leq c_{max}.$$

Thus,

$$D \leq c_{max} \leq d_{max}.$$

This gives an effective bound on the maximum absolute deviation c_{max}. So, if there is a considerable difference between D and d_{max}, it might be better to apply the Chebyshev criterion.

3 Obtaining a Least-Squares Fit (or Regression)

Linear Regression

Suppose a model of the form $y = ax + b$ is expected, and it has been decided to use the m data points $(x_i, y_i), i = 1, 2, \ldots, m$ to estimate a and b. Applying the least-squares criterion in equation (1), we find that we must minimize

$$E = \sum_{i=1}^{m} [y_i - f(x_i)]^2 = \sum_{i=1}^{m} (y_i - ax_i - b)^2 .$$

In order to minimize E, we need to compute and set to 0 the partial derivatives of E with respect to both a and b:

$$\frac{\partial E}{\partial a} = -2\sum_{i=1}^{m}(y_i - ax_i - b)\,x_i = 0$$
$$\frac{\partial E}{\partial b} = -2\sum_{i=1}^{m}(y_i - ax_i - b) = 0$$

Rewriting these equations, we obtain the following system of two equations in two unknowns

$$a\sum_{i=1}^{m}x_i^2 + b\sum_{i=1}^{m}x_i = \sum_{i=1}^{m}x_iy_i$$
$$a\sum_{i=1}^{m}x_i + mb = \sum_{i=1}^{m}y_i,$$

which may easily be solved to yield

$$a = \frac{m\sum x_iy_i - \sum x_i\sum y_i}{m\sum x_i^2 - \left(\sum x_i\right)^2} \tag{2}$$

and

$$b = \frac{\sum x_i^2\sum y_i - \sum x_iy_i\sum x_i}{m\sum x_i^2 - \left(\sum x_i\right)^2} \tag{3}$$

Maple can be used to determine the linear least-squares fit for any set of data.

Example: The following table shows the average age versus height for children. The data is copied from *An Introduction to the Mathematics of Biology, with Computer Algebra Models* by Yeargers, Shonkwiler, and Herod. A plot of the data (height vs. age) reveals that the

height (cm)	75	92	108	121	130	142	155
age	1	3	5	7	9	11	13

data appears to be linear in nature, so we will perform a linear regression to determine the least-squares best linear fit, letting x_i represent the age and y_i represent the height. Working out the necessary sums, we find that

$$\sum x_i = 49$$
$$\sum x_i^2 = 455$$
$$\sum y_i = 823$$
$$\sum x_iy_i = 6485.$$

Applying the formulas in (2) and (3), we obtain $a = 6.4643$ and $b = 72.3214$. Therefore, the data is approximated by the curve $y = 6.4643x + 72.3214$. See also the Maple example *regression.mw* under Linear Regression.

4 Plotting the Residuals for a Least Squares Fit

We plot the residuals in the y-axis and the independent variable on the x-axis. The deviations should be randomly distributed and contained in a reasonably small band that is consistent with the accuracy required by the model. If the residual is excessively large, this means that the data point in question should be examined to discover the cause of the large deviation (i.e., did we miss an outlier or is something else going on). A pattern or trend in the residuals indicates that a predictable effect remains to be modeled, and the nature of hte pattern gives a hint as to how to refine the model. If the residual is increasing in absolute value, then that is an indication that our model is unacceptable.

We will use a Maple example to illustrate this.

5 Fitting a Power Curve

Suppose we think that the curve will be of the form $y = Ax^N$, where we know what N is (usually unlikely). Applying the least-squares method to fit the curve to a set of m data points requires minimization of

$$E = \sum_{i=1}^{m} [y_i - f(x_i)]^2 = \sum_{i=1}^{m} \left(y_i - Ax_i^N\right)^2.$$

A necessary condition for optimality is that the derivative of E with respect to A equals 0, giving

$$\frac{dE}{dA} = -2 \sum_{i=1}^{m} x_i^N \left(y_i - Ax_i^n\right) = 0,$$

sor

$$A = \frac{\sum x_i^N y_i}{\sum x_i^{2N}}.$$

However, it is not likely that we "know" the value of N in advance. What if we do not know N? In other words, now assume that we seek a least-squares estimate of the form $y = ax^n$, where both a and n are parameters to be determined. In this case, we will first need to transform the function by taking the logarithm of both sides to obtain

$$\ln y = \ln a + n \ln x. \tag{4}$$

Then, we simply transform the data by taking the natural log of both x_i and y_i, then apply linear regression to obtain $\ln a$ and n. The parameter a is then found by $a = e^{\ln a}$.

Example: The following table shows the ideal weights for medium built males. The data is copied from *An Introduction to the Mathematics of Biology, with Computer Algebra Models* by Yeargers, Shonkwiler, and Herod.

This example is worked out on the Maple worksheet, *regression.mw*.

Height (in)	Weight (lb)
62	128
63	131
64	135
65	139
66	142
67	146
68	150
69	154
70	158
71	162
72	167
73	172
74	177

First, we try a linear least-squares fit. As we can see from the plot of the residuals, the residual is increasing as the height increases. Therefore, this model does not seem to be acceptable. However, since we know that in many geometric solids, the volume changes with the cube of the height, we will try to find a cubic fit for the data. The results from the cubic fit are much better: the residual is not large, and it decreases as the height increases. We would like to know if we can do better. So, we will try to find a fit of the form

$$\text{weight } = A(\text{height } - 60)^n.$$

Ideally, n is an integer. However, we discover that $n \approx 0.168$. Since the mathematics behind our approach does not require n to be an integer, this is okay. If we wish to make n a rational number, then the nearest rational number to the value of n we obtained is $\frac{1}{6}$. So, we also look at the results with this value of n.

6 Multiple Regression

In many real problems, there are more than two variables to be considered. For example,

(1) The pressure of a gas depends on the volume V and the temperature T: $PV = kT$.

(2) The relationship between the surface area of a human as a function of height and weight:
$$\text{SA} = c(\text{wt})^{\alpha}(\text{ht})^{\beta}.$$

(3) Determination of percent body fat from body mass index and skin fold:
$$\%\text{ body fat} = a(\text{BMI}) + b(\text{skinfold}) + c.$$

Look at (2), for example. Taking the natural log of both sides gives

$$\ln(\mathrm{SA}) = a\ln(\mathrm{wt}) + b\ln(\mathrm{ht}) + \ln c,$$

so we obtain an expression similar to that in (3).

In general, we have a relation of the form

$$Y = a_1X_1 + a_2X_2 + \cdots + a_rX_r.$$

Suppose that we have n data points, $(X_{1,i}, X_{2,i}, \ldots, X_{r,i}, Y_i)\,, i = 1, 2, \ldots, n$. Our goal is to minimize the least-squared error:

$$E = \sum_{i=1}^{n} \left(Y_i - (a_1X_{1,i} + a_2X_{2,i} + \cdots + a_rX_{r,i})\right)^2$$

Differentiating with respect to each parameter a_i and setting the derivatives to zero gives the following system of equations

$$\begin{aligned} a_1 \sum_{i=1}^{n} X_{1,i}X_{1,i} + \cdots + a_r \sum_{i=1}^{n} X_{1,i}X_{r,i} &= \sum_{i=1}^{n} X_{1,i}Y_i \\ &\vdots \\ a_1 \sum_{i=1}^{n} X_{r,i}X_{1,i} + \cdots + a_r \sum_{i=1}^{n} X_{r,i}X_{r,i} &= \sum_{i=1}^{n} X_{r,i}Y_i. \end{aligned}$$

Solve by writing the equations in matrix form. This is easily done in various computer algebra systems, such as Maple.

Maple code for examples is in the file *regression.mw*.