Model Fitting

Institute of Technology of Cambodia

Departerment of Applied Mathematics and Statistics.

Tepmony SIM & **Sokea LUEY**

November , 2022

Course outline

Introduction

The preceding discussion identifies three possible tasks when we are analyzing a collection of data points:

1. Fitting a selected model types to the data.
2. Choosing the most appropriate model from competing types that have been fitted. **For Example,** we may need to determine whether the best-fitting exponential model is a better model than the best fitting polynomial model.
3. Making predictions from the collected data.

Introduction

Relationship Between Model Fitting and Interpolation

1. There are some deviation between the model and the collected data points to have a model that satisfactorily explains the situation under investigation
2. Both model and collected data possibly contain errors
3. **Model Fitting:** we except some scatter in the experimental data, we want the best model of a given form that can explain the data (explicative significance) - "theory driven"
4. **Model Interpolation:** less explicative significance - "data driven", focus mainly on predictive performance

Today we'll be focussing on **Model Fitting**

Introduction

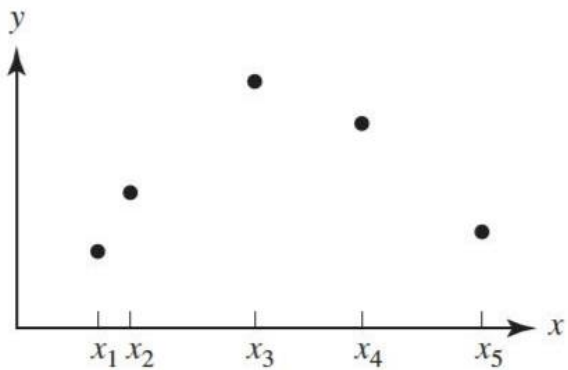


Figure: Observations relating the variable y and x

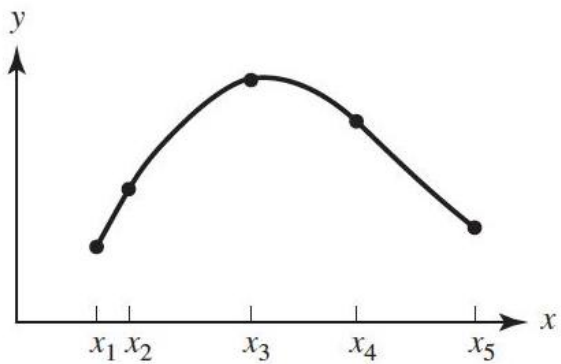


Figure: interpolating the data using a smooth polynomial

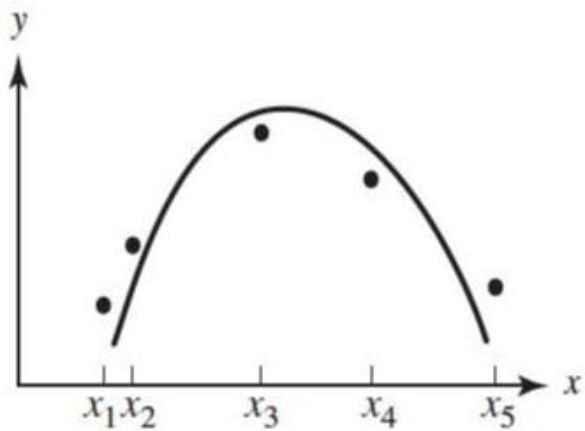


Figure: Fitting a parabola $y = c_1x^2 + C_2x + C_3$

Introduction

Sources of Error in the Modeling Process

There are 4 Sources and types of Error

1. **Formulation error:**These errors result from the assumption that certain variables are negligible or from simplifications in describing interrelationships among the variables in the various submodels.
2. **Truncation error:** 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.

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$$

3. **Round-off error:** These errors are caused by using a finite digit machines for computations. **For Example,** consider a calculator or computer that uses 8-digit arithmetic.

Introduction

Sources of Error in the Modeling Process Cont.

$\frac{1}{3}$ is 0.33333333 so that 3 times $\frac{1}{3}$ is 0.99999999 rather than the actual value 1. The error 10^{-8} is due to round-off. The real number $\frac{1}{3}$ is an infinite string of decimal digits $.3333\cdots$. but any calculator or computer can do arithmetic only with number having finite precision. Round-off is just one of the things we have to live with *-and be aware of-* when we use computing machines.

4. **Measurement error:** 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.

Fitting Models to Data Graphically

Fitting Models to Data Graphically

If the experiment has been carefully designed and the trial meticulously conducted, the modeler needs to appraise the accuracy of the data before attempting to fit the model. How were the data collected? What is the accuracy of the measuring device use in the collected process? Do any points appear suspicious?

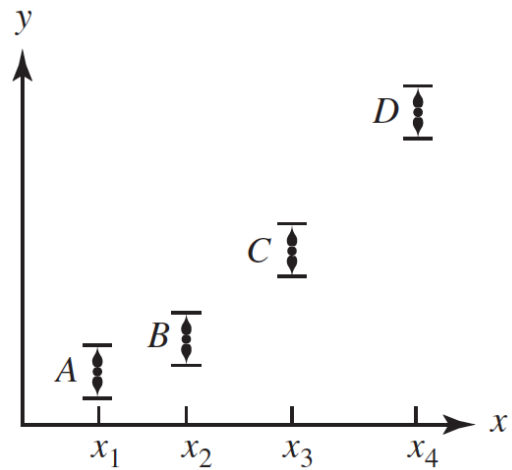


Figure: Each data point is thought of as an interval of confidence.

Fitting Models to Data Graphically Cont.

Visual Model Fitting with the Original Data

Suppose we want to fit the model $y = ax + b, a$ and b are constant to the data show in **Figure 3.4**. How might we choose a & b to determind the line that best fit the data?

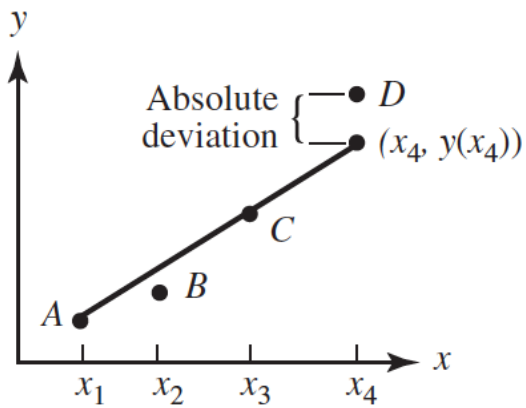


Figure: 3.5: Each data point is thought of as an interval of confidence.

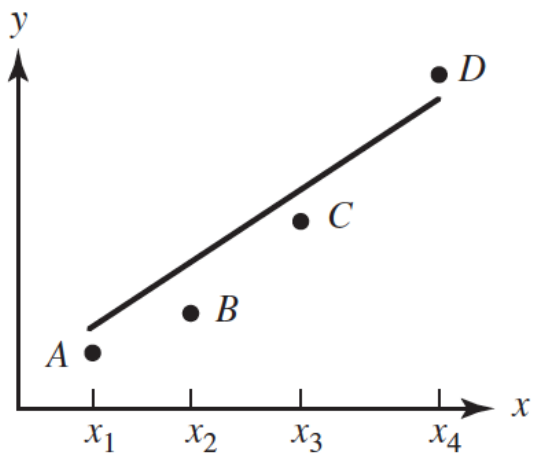


Figure: 3.6: Minimizing the largest absolute deviation from the fitted line.

Fitting Models to Data Graphically Cont.

Transforming the Data

Suppose that $y = Ce^x$ is suspected for some submodel and the data show in Table 1. How can we graphically as models?

Table 3.1 Collected data

x	1	2	3	4
y	8.1	22.1	60.1	165

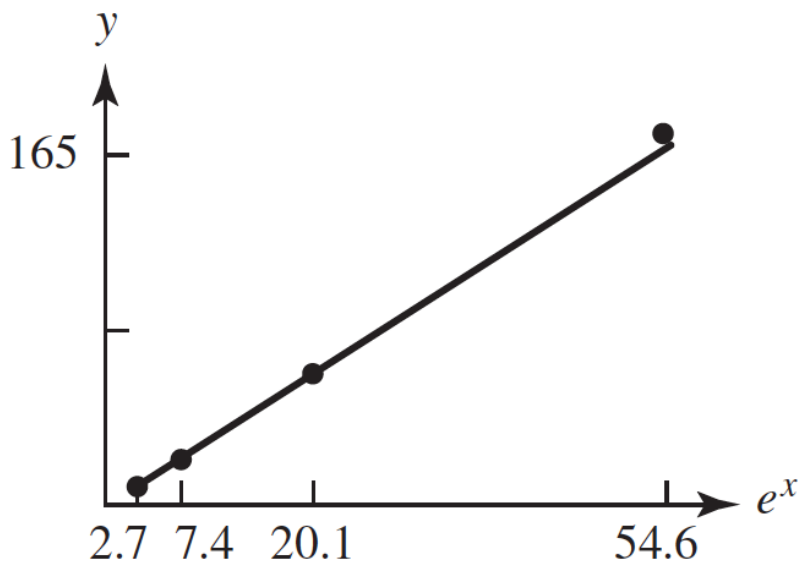


Figure: 3.7: Plot of y versus e^x for the data given in Table 3.1

Fitting Models to Data Graphically Cont.

Example 1

Now let's consider an alternative technique that is useful in a variety of problems. Take the logarithm of each side of the equation $y = Ce^x$.

Table 3.2 The transformed data from Table 3.1

x	1	2	3	4
$\ln y$	2.1	3.1	4.1	5.1

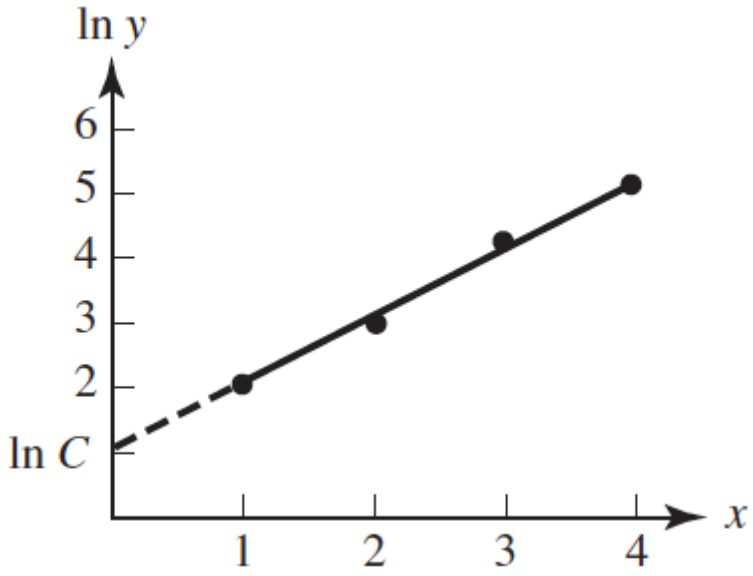


Figure: 3.8: Plot of $\ln y$ versus x using Table 3.2

Analytic Methods of Model Fitting

Chebyshev Approximation Criterion

Let's analyze this geometric construction. Given a collection of m data points $(x_i, y_i), i = 1, 2, ..., m$, fit the collection to the line $y = ax + b$, determined by parameters a and b, that minimizes the distance between (x_i, y_i) and its corresponding data point on the line $(x_i, ax_i + b)$. That is, minimize the largest absolute deviation $|y_i - y(x_i)|$ over the entire collection of data points. Now let's generalize this criterion.

Given some function type $y = f(x)$ and a collection of m data points (x_i, y_i), minimize the largest absolute deviation $|y_i - f(x_i)|$ over the entire collection. That is, determine the parameters of the function type $y = f(x)$ that minimizes the number

$$\text{Maximum } |y_i - f(x_i)| \qquad i = 1, 2, ..., m \tag{3.1}$$

This criterion is called **Chebyshev Approximation Criterion.**

Analytic Methods of Model Fitting Cont.

Minimizing the Sum of the Absolute Deviations

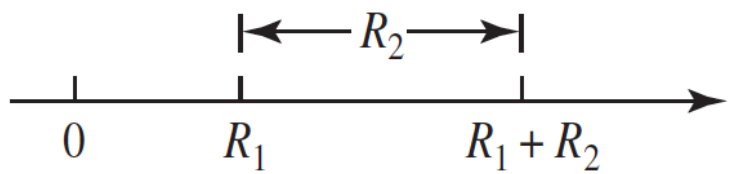
Given some function type $y = f(x)$ and a collection of m data points $(x_i, y_i), bi = 1, 2, ..., m$, we week to minimize

$$\sum_{i=1}^{m} |y_i - f(x_i)| \tag{3.2}$$

If we let $R_i = |y_i - f(x_i)|, i = 1, 2, ..., m$ be each absolute deviation, then 3.2 can be interpreted as minimizing the length of the line formed by adding together the number R_i. This is illustrated for $m = 2$ in Figure 3.13.

Figure 3.13

A geometric interpretation of minimizing the sum of the absolute deviations



Analytic Methods of Model Fitting Cont.

Least-Squares Criterion

Currently, the most frequently used curve-fitting criterion is the **least-squares criterion.** If we use the same notation shown earlier, the problem is to determine the parameters of the function type $y = f(x)$ to minimize the sum

$$\sum_{i=1}^{m} |y_i - f(x_i)|^2 \tag{3.3}$$

Consider the case of three data points and let $R_i = |y_i - f(x_i)|$ denote the absolute deviation between the observed and predicted values for$i = 1, 2, 3$. Think of the R_i as the scalar components of a deviation vector, as depicted in Figure 3.14. Thus the vector $R = R_1 i + R_2 j + R_3 k$ represents the resultant deviation between the observed and predicted values.

Analytic Methods of Model Fitting Cont.

Least-Squares Criterion

The magnitude of the deviation vector is given by

$$|R|^2 = \sqrt{R_1^2 + R_2^2 + R_3^2}$$

o minimize $|R|$ we can minimize $|R|^2$ (see Problem 1). Thus, the last-squares problem is to determine the parameters of the function type $y = f(x)$ such that

$$|R|^2 = \sum_{i=1}^{3} R_i^3 = \sum_{i=1}^{3} |y_i - f(x_i)|^2$$

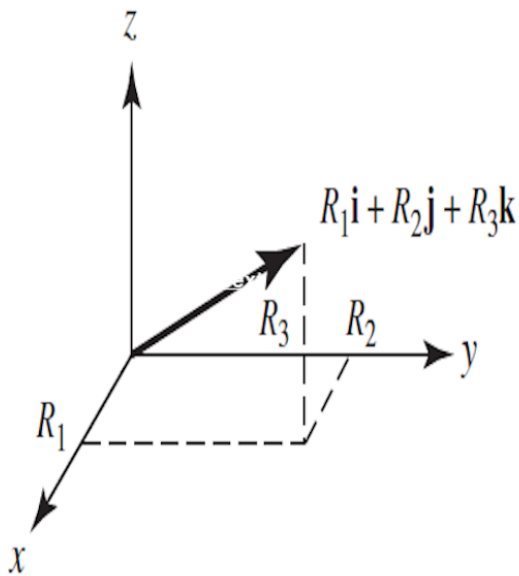


Figure: 3.14A geometric interpretation of the least-squares criteria

Analytic Methods of Model Fitting Cont.

Relating the Criteria

Suppose the Chebyshev criterion is applied and the resulting optimization problem solve to yield the function $f_1(x)$. The absolute deviations resulting from the fit are defined as followings:

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

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 us the minimal largest absolute deviation obtainable.
Suppose the least-squares criterion is applied and the resulting optimization problem solved to yield the function $f_2(x)$. The absolute deviations resulting from the fit are then given by

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

Analytic Methods of Moedel Fitting Cont.

Relating the Criteria Cont.

Define d_{max} as the largest of the absolute deviations d_i. Clearly, $d_{max} \geq c_{max}$. Since their 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 $i = 1, 2, \cdots, m$. 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,

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

Analytic Methods of Moedel Fitting Cont.

Relating the Criteria Cont.

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

Applying the Lest-Squares Criterion

Fitting a Straight Line

Suppose a model of the form $y = Ax + B$ is expected and it has been decided to use the m data points $(x_i, y_i),\ i = 1, 2, \cdots, m.$ to estimate A and B. Denote the least-squares criterion 3.3 to this situation requires the minimization of

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

A necessary condition for optimality is that the two partial l derivatives $\partial S/\partial a$ and $\partial S/\partial b$ equal zero, yielding the equations

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

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

Applying the Lest-Squares Criterion Cont.

Fitting a Straight Line Cont.

These equations can be rewritten to give

$$\begin{cases} a + \sum_{i=1}^{m} x_i^2 + b \sum_{i=1}^{m} x_i & = \sum_{i=1}^{m} x_i y_i \\ a \sum_{i=1}^{m} x_i + mb & = \sum_{i=1}^{m} y_i \end{cases} \tag{3.4}$$

which may easily be solve to yield

$$a = \frac{m \sum x_i y_i - \sum x_i y_i}{m \sum x_i^2 - (\sum x_i)^2}, \quad \text{the } \textbf{slope} \tag{3.5}$$

$$b = \frac{\sum x_i^2 \sum y_i - \sum x_i y_i \sum x_i}{m \sum x_i^2 - (\sum x_i)^2} \quad \text{the } \textbf{intercept} \tag{3.6}$$

Computer codes are easily written to compute these values for a and b for any collection of data points. Equation (3.4) are called the **normal equations.**

Applying the Lest-Squares Criterion Cont.

Fitting a Power Curve

Let's use the least-squares criterion to fit a curve of $y = Ax^n, n$ is fixed, to a given collection of data points. Call the least-squares estimate of the model $f(x) = ax^n$. Application of the criterion then requires minimization of

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

A necessary condition for optimality is $dS/da = 0$, giving the equation

$$\frac{dS}{da} = -2 \sum_{i=1}^{m} x_i^n [y_i - ax_i^n] = 0$$

Solving the equation for a yields

$$a = \frac{\sum x_i^n y_i}{\sum x_i^{2n}}, \qquad n \text{ is } fixed, \tag{3.7}$$

Appying the Lest-Squares Criterion Cont.

Fitting a Power Curve Cont.

Example: Let's fit $y = Ax^2$ to data shown in Table 3.3 and predict the value of when $x = 2.25$.

Teble 3.3 Data collected to fit $y = Ax^2$

x	0.5	1.0	1.5	2.0	2.5
y	0.7	3.4	7.2	12.4	20.1

Applying the Lest-Squares Criterion Cont.

Transformed Least-Squares Fit

Example: consider fitting the model $y = Ae^{Bx}$ using the least-square criterion. Call the least-squares estimate of the model $f(x) = ae^{bx}$. Application of the criterion then requires the minimization of

$$S = \sum_{i=1}^{m} [y_i - f(x_i)]^2 = \sum_{i=1}^{m} [y_i - ae^{bx_i}]^2$$

A necesary condition $\partial S/\partial a = \partial S/\partial b = 0$

Applying the Lest-Squares Criterion Cont.

Transformed Least-Squares Fit

Taking the logarithm of both sides of the equation $y = \alpha x^n$ yield

$$\ln y = \ln \alpha + n \ln x \tag{3.8}$$

(3.8) is straight line when $\ln y$ and $\ln x$ are variable.
Using (3.5) and (3.6) to solve for n and $\ln \alpha$ with transformed variables and $m = 5$ data points, we obtain

$$n = \frac{5\sum(\ln x_i)(\ln y_i) - (\sum \ln x_i)(\sum \ln y_i)}{5\sum(\ln x_i)^2 - (\sum \ln x_i)^2}$$

$$\ln \alpha = \frac{\sum(\ln x_i)(\ln y_i) - (\sum \ln x_i \ln y_i)(\sum \ln x_i)}{5\sum(\ln x_i)^2 - (\sum \ln x_i)^2}$$

Using data displayed in Table 3.3 to calculate the least-squares best fit of (3.8).

Choosing a Best Model

Table 3.4 Deviations between the data in Table 3.3 and the fitting model $y = 3.1869x^2$

x_i	0.5	1.0	1.5	2.0	2.5
y_i	0.7	3.4	7.2	12.4	20.1
$y_i - y(x_i)$	-0.0967	0.2131	0.02998	-0.3476	0.181875

Since $y_i - y(x_i) = d_i, i = 1, 2, 3, 4, 5.$ is deviation, so we obtain:

$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 = 0.20954$$

So, $D = (0.20954/5)^2 = 0.204714.$
Since $|d_i| \leq 0.3476$, when $x = 2.0$ we obtain

$$D = 0.204714 \leq c_{max} \leq 0.3476 = d_{max}$$

Let's find c_{max}. Since there are five data points, the mathematical problem is to minimize the largest of the five number

$$|r_i| = |y_i - y(x_i)|, i = 1, 2, 3, 4, 5.$$

Choosing a Best Model Cont.

Let r be the largest number, we will minimize r subject to $r \geq r_i$ and $r \geq -r_i$. Denote our model by $y(x) = a_2 x^2$. By using the data in Table 3.3 we obtain the following linear program:

Minimize r subject to:

$$r - r_1 = r - (0.7 - 0.25a_2) \geq 0$$
$$r + r_1 = r + (0.7 + 0.25a_2) \geq 0$$
$$r - r_2 = r - (3.4 - a_2) \geq 0$$
$$r + r_2 = r + (3.4 + a_2) \geq 0$$
$$r - r_3 = r - (7.2 - 2.25a_2) \geq 0$$
$$r + r_3 = r + (7.2 - 2.25a_2) \geq 0$$
$$r - r_4 = r - (12.4 - 4a_2) \geq 0$$
$$r + r_4 = r + (12.4 - 4a_2) \geq 0$$
$$r - r_5 = r - (20.1 - 6.25a_2) \geq 0$$
$$r + r_5 = r + (20.1 + 6.25a_2) \geq 0$$

Choosing a Best Model Cont.

In Chapter 7 we show that the solution of the preceding linear program yields $r = 0.28293$ and $a/2 = 3.17073$. Thus, we have reduced our largest deviation from $d_{max} = 0.3476$ to $c_{max} = 0.28293$. Note that we can reduce the largest deviation no further than 0.28293 for the model type $y = Ax^2$.