Chapter 4
Curve Fitting: Least-Squares Approximation

Dr. Kareem Elgindy

Lecturer
Mathematics Department, Faculty of Science
Assiut University

16 November, 2015

Outline

Outline

Introduction

- There are two general approaches for **curve fitting** that are distinguished from each other **on the basis of the amount of error associated with the data**:
  1. When the **data** are known to be **very precise**, the basic approach is ***interpolation***; i.e. to fit a curve that passes directly through each of the points. Such data usually originate from tables.

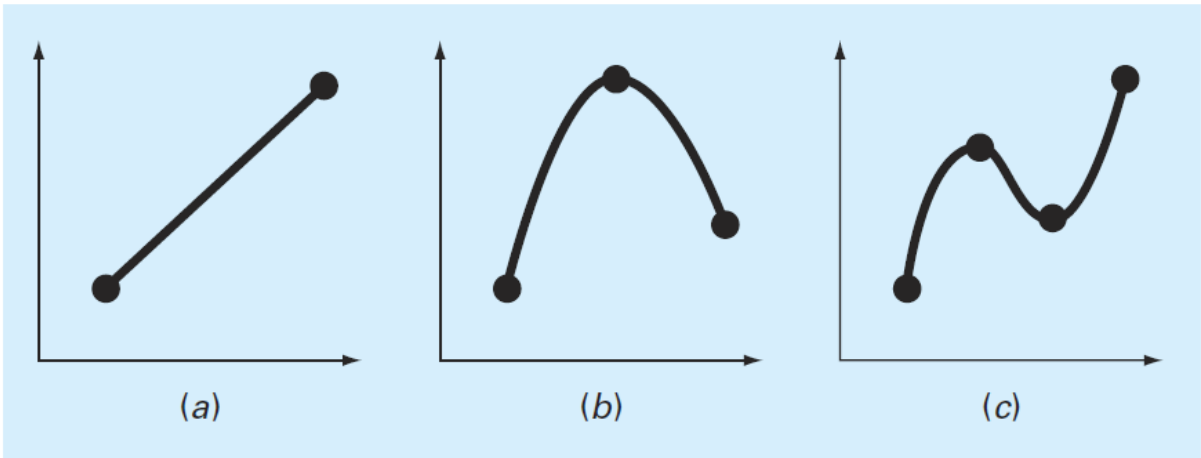

Figure 1: Three attempts to fit "best" curves through two, three, and four data points via linear, quadratic, and cubic interpolation.

Introduction

- There are two general approaches for **curve fitting** that are distinguished from each other **on the basis of the amount of error associated with the data**:
  2. When the **data** exhibit a **significant degree of error** or **"scatter,"** the strategy is to ***derive a single curve that represents the general trend/pattern of the data taken as a group***.

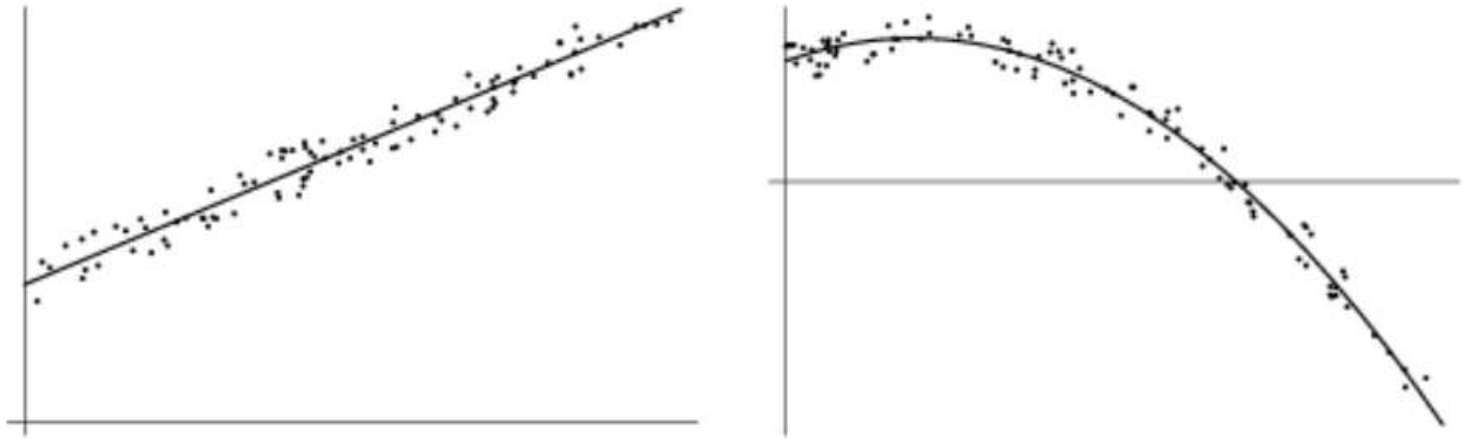

Figure 2: Two attempts to fit the "best" straight line and parabola through two sets of scattered data points.

Introduction

- Interpolation is not suitable in the latter case, because any individual data point may be incorrect, so **we make no effort to intersect every point**.
  - In other words, it is ***unreasonable to require that the approximating function agree exactly with the data***, and it would seem more reasonable to **find the curve that best approximates all the data points (in some sense)**. One approach of this nature is called ***least-squares approximation (least-squares regression)***.

Outline

Linear Least-Squares Regression

Suppose that we have m data points $\{x_i, y_i\}_{i=1}^m$ that exhibit a certain degree of error, where $\{x_i\}_{i=1}^m$ are the values of the independent variable, and $\{y_i\}_{i=1}^m$ are the values of the dependent variable (**the observed values**). Also assume that $P(x_i) = a_1 x_i + a_0, i = 1, \ldots, m$, are the corresponding approximating values (**the predicted values**).

- The problem of finding the equation of the best linear approximation in the absolute sense requires that values of a_0 and a_1 be found to minimize the maximum absolute error, $E_\infty(a_0, a_1)$, defined by

$$E_\infty(a_0, a_1) = \max_{1 \le i \le m} |y_i - P(x_i)| = \max_{1 \le i \le m} |y_i - (a_1 x_i + a_0)| . \tag{1}$$

  This is commonly called a **minimax problem**, and unfortunately it **cannot be handled by elementary techniques**.

Linear Least-Squares Regression

- Another approach to determine the best linear approximation involves finding the values of a_0 and a_1 that minimize

$$E_1(a_0, a_1) = \sum_{i=1}^{m} |y_i - P(x_i)| = \sum_{i=1}^{m} |y_i - (a_1 x_i + a_0)|. \tag{2}$$

This quantity is called the **absolute deviation**, and it represents **the sum of the differences between the observed values $\{y_i\}_{i=1}^m$ and the predicted values $\{P(x_i)\}_{i=1}^m$**.

Linear Least-Squares Regression

- To minimize a function of two variables, we need to set its partial derivatives to zero and simultaneously solve the resulting equations. In the case of the absolute deviation, we need to find a_0 and a_1 with

$$0 = \frac{\partial}{\partial a_0} E_1(a_0, a_1) = \frac{\partial}{\partial a_0} \sum_{i=1}^{m} |y_i - (a_1 x_i + a_0)|; \quad \text{(3a)}$$

$$0 = \frac{\partial}{\partial a_1} E_1(a_0, a_1) = \frac{\partial}{\partial a_1} \sum_{i=1}^{m} |y_i - (a_1 x_i + a_0)|. \quad \text{(3b)}$$

The problem with this approach is that the **absolute-value function is not differentiable at zero**, and we might not be able to find solutions to this pair of equations.

Linear Least-Squares Regression

The **least-squares approach**[1] to this problem involves determining the best approximating line when the error involved is **the sum of the squares of the differences between the observed values $\{y_i\}_{i=1}^m$ and the predicted values $\{P(x_i)\}_{i=1}^m$**. Hence, the constants a_0 and a_1 must be found to minimize the least-squares error:

$$E_2(a_0, a_1) = \sum_{i=1}^{m} [y_i - P(x_i)]^2 = \sum_{i=1}^{m} [y_i - (a_1 x_i + a_0)]^2. \tag{4}$$

The least-squares method is the most convenient procedure for determining best linear approximations.

[1] In 1805 the French mathematician **Adrien-Marie Legendre** published the first known recommendation to use the line that minimizes the sum of the squares of errors (deviations); i.e., the *modern least-squares approximation*.

Linear Least-Squares Regression

Definition 1 (**Least-Squares Method**)

The method of least-squares is used to **estimate parameters** in mathematical models **by minimizing the sum of the squares of the differences between the observed values and the predicted values** under the model.

Linear Least-Squares Regression

For a minimum to occur, we need

$$0 = \frac{\partial}{\partial a_0} E_2(a_0, a_1) = \frac{\partial}{\partial a_0} \sum_{i=1}^{m} [y_i - (a_1 x_i + a_0)]^2$$

$$= 2 \sum_{i=1}^{m} (y_i - a_1 x_i - a_0)(-1); \tag{5a}$$

$$0 = \frac{\partial}{\partial a_1} E_2(a_0, a_1) = \frac{\partial}{\partial a_1} \sum_{i=1}^{m} [y_i - (a_1 x_i + a_0)]^2$$

$$= 2 \sum_{i=1}^{m} (y_i - a_1 x_i - a_0)(-x_i). \tag{5b}$$

Linear Least-Squares Regression

These equations simplify to the **normal equations**:

$$a_0 \cdot m + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i; \tag{6a}$$

$$a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i. \tag{6b}$$

Linear Least-Squares Regression

The unique solution to this system of equations is

$$a_0 = \frac{\sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i - \sum_{i=1}^{m} x_i y_i \sum_{i=1}^{m} x_i}{m\left(\sum_{i=1}^{m} x_i^2\right) - \left(\sum_{i=1}^{m} x_i\right)^2}; \tag{7a}$$

$$a_1 = \frac{m\sum_{i=1}^{m} x_i y_i - \sum_{i=1}^{m} x_i \sum_{i=1}^{m} y_i}{m\left(\sum_{i=1}^{m} x_i^2\right) - \left(\sum_{i=1}^{m} x_i\right)^2}. \tag{7b}$$

Linear Least-Squares Regression

Example 1

Find the least-squares line approximating the data in the table below.

x_i	y_i	x_i	y_i
1	1.3	6	8.8
2	3.5	7	10.1
3	4.2	8	12.5
4	5.0	9	13.0
5	7.0	10	15.6

Linear Least-Squares Regression

Solution 1

We first extend the table to include x_i^2 and $x_i y_i$, and sum the columns as shown in the table below.

x_i	y_i	x_i^2	$x_i y_i$
1	1.3	1	1.3
2	3.5	4	7.0
3	4.2	9	12.6
4	5.0	16	20.0
5	7.0	25	35.0
6	8.8	36	52.8
7	10.1	49	70.7
8	12.5	64	100.0
9	13.0	81	117.0
10	15.6	100	156.0
55	81.0	385	572.4

Linear Least-Squares Regression

Solution 1

The solution of the normal equations is given by Equations (7) *as follows:*

$$a_0 = \frac{385(81) - 55(572.4)}{10(385) - (55)^2} = -0.360; \tag{8a}$$

$$a_1 = \frac{10(572.4) - 55(81)}{10(385) - (55)^2} \approx 1.538. \tag{8b}$$

So $P(x) \approx 1.538\,x - 0.360$.

Linear Least-Squares Regression

Solution 1

The graph of this line and the data points are shown in the figure below.