MA 138 – Calculus 2 with Life Science Applications
Curve Fitting — Least Squares Approximation
(Handout)

Alberto Corso
⟨alberto.corso@uky.edu⟩

Department of Mathematics
University of Kentucky

March 24 & 27, 2017

Data Analysis and Curve Fitting

Imagine that we are studying a physical system involving two quantities:
$$x \quad \text{and} \quad y.$$

Also suppose that we expect a linear relationship between these two quantities, that is, we expect $y = ax + b$, for some constants a and b.

We wish to conduct an experiment to determine the value of the constants a and b. We collect some data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, which we plot in a rectangular coordinate system.

Since we expect a linear relationships, all these points should lie on a single straight line: The slope of this line will be a, and the intercept is b.

In other words, we should have that the following system of linear equations has exactly one solution

$$\begin{cases} ax_1 + b &= y_1 \\ ax_2 + b &= y_2 \\ & \vdots \\ ax_n + b &= y_n \end{cases} \quad \rightsquigarrow \quad \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

that is, we should expect the system of linear equations above to be consistent.

Unfortunately, when we plot our data, we discover that our points do not lie on a single line. This is only to be expected, since our measurements are subject to experimental error.

On the other hand, it appears that the points are approximately "collinear."

It is our aim to find a straight line with equation

$$y = \widehat{a}\,x + \widehat{b}$$

which fits the data "*best*."

Of course, optimality could be defined in many different ways.

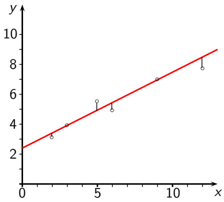

FIGURE 1: Fitting a straight line to data by the method of least squares

It is customary to proceed as follows. Consider the *deviations* (differences)

$$\delta_1 = (ax_1 + b) - y_1, \quad \delta_2 = (ax_2 + b) - y_2, \quad \ldots, \quad \delta_n = (ax_n + b) - y_n.$$

If all the data points were to be lying on a straight line then there would be a unique choice for a and b such that all the deviations are zero.

In general they aren't. Which of the deviations are positive, negative or exactly zero depends on the choice of the parameters a and b.

As a condition of optimality we minimize the square root of the sum of the squares of the deviations ("least squares"), that is, we choose $\widehat{a}$ and $\widehat{b}$ in such a way that $\sqrt{\delta_1^2 + \delta_2^2 + \ldots + \delta_n^2}$ is as small as possible.

Remark

This kind of analysis of data is also called *regression analysis*, as one of the early applications of least squares was to genetics, to study the well-known phenomenon that children of unusually tall or unusually short parents tend to be more normal in height than their parents. In more technical language, the children's height tends to "regress toward the mean."

If you have taken a Statistics class, you might have seen these formulas

$$\widehat{a} = \frac{n\left(\sum_{i=1}^n x_i y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2} \qquad \widehat{b} = \frac{1}{n}\left(\sum_{i=1}^n y_i - a\sum_{i=1}^n x_i\right),$$

which give the optimal solution to our least squares approximation problem. There is *no need to memorize* these formulas. The discussion that follows will explain how these formulas are obtained!

Remark/Example

Suppose we are looking for a linear relationship between more than two quantities!

For example, a consulting firm has been hired by a large SEC university to help make admissions decisions.

Each applicant provides the university with three pieces of information: their score on the SAT exam, their score on the ACT exam, and their high school GPA (0-4 scale).

The university wishes to know what weight to put on each of these numbers in order to predict student success in college.

The consulting firm begins by collecting data from the previous year's freshman class. In addition to the admissions data, the firm collects the student's current (college) GPA (0-4 scale), say C_GPA.

A partial listing of the firm data might look like

SAT	ACT	GPA	C_GPA
600	30	3.0	3.2
500	28	2.9	3.0
750	35	3.9	3.5
650	30	3.5	3.5
550	25	2.8	3.2
800	35	3.7	3.7
⋮	⋮	⋮	⋮

Ideally, the firm would like to find numbers (weights) x_1, x_2, x_3 such that for all students

$$(\text{SAT})x_1 + (\text{ACT})x_2 + (\text{GPA})x_3 = (\text{C\_GPA}).$$

These numbers would tell the university exactly what weight to put on each piece of data. Statistically, of course, it is highly unlikely that such numbers exist. Still, we would like to have an approximate "best" solution.

Remark/Example

Instead of a linear relationship among the two quantities x and y involved in our original physical system, suppose that we expect a quadratic relationship. That is, we expect

$$y = ax^2 + bx + c,$$

for some constants a, b, and c. This means that all our plotted data points should lie on a single parabola. In other words, the system of equations below should have exactly one solution

$$\left\{\begin{array}{rcl} ax_1^2 + bx_1 + c & = & y_1 \\ ax_2^2 + bx_2 + c & = & y_2 \\ & \vdots & \\ ax_n^2 + bx_n + c & = & y_n \end{array}\right. \quad \rightsquigarrow \quad \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

or, in other words, the system of linear equations should be consistent.

Again, this is unlikely since data measurements are subject to experimental error. As we mentioned, if the exact solution does not exists, we seek to find the equation of the parabola $y = \widehat{a}\,x^2 + \widehat{b}\,x + \widehat{c}$ which fits our given data best.

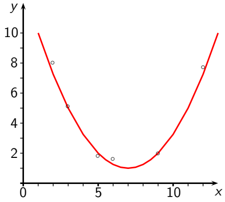

FIGURE 2: Fitting a parabola to data by the method of least squares

General problem

In our all previous examples, our problem reduces to finding a solution to a system of n linear equations in m variables, with $n > m$. Using our traditional notations for systems of linear equations, we translate our problem into matrix notation. Thus, we are seeking to solve

$$A\,\mathbf{x} = \mathbf{b},$$

where A is an $n \times m$ given matrix (with $n > m$), $\mathbf{x}$ is a column vector with m variables, and $\mathbf{b}$ is a column vector with n given entries.

Example 1

Find a solution to

$$\begin{bmatrix} -1 & 2 \\ 2 & -3 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 1 \\ 2 \end{bmatrix}.$$

"Best" approximate solution to our general problem

Now, instead of looking for a solution to our given system of linear equations (which, in general, we don't have!) we could look for an approximate solution. To this end, we recall that for a given vector $\mathbf{v} = [v_1, v_2, \dots, v_n]^T$ its length is defined to be

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \dots + v_n^2}.$$

(This is a generalized version of Pythagoras' Theorem!)

If A is an $n \times m$ matrix, $\mathbf{x}$ is a column vector with m entries and $\mathbf{b}$ is a column vector with n entries, a *least squares solution* to the equation $A\mathbf{x} = \mathbf{b}$ is a vector $\widehat{\mathbf{x}}$ so that the length of the vector $A\widehat{\mathbf{x}} - \mathbf{b}$, that is $\|A\widehat{\mathbf{x}} - \mathbf{b}\|$, is as small as possible. In other words

$$\|A\widehat{\mathbf{x}} - \mathbf{b}\| \le \|A\mathbf{z} - \mathbf{b}\|$$

for every other vector $\mathbf{z}$.

Least Squares Solution

How do we find this? This is answered in the following Theorem.

Theorem

The least squares solution $\widehat{\mathbf{x}}$ to the system of linear equations $A\mathbf{x} = \mathbf{b}$, where A is an $n \times m$ matrix with $n > m$, is a/the solution $\widehat{\mathbf{x}}$ to the associated system (of m linear equations in m variables)

$$(A^T A)\,\mathbf{x} = A^T\mathbf{b},$$

where A^T denotes the transpose matrix of A.

Note: the matrix $A^T A$ in the Theorem is a symmetric, square matrix of size $m \times m$. If it is invertible, we can then expect exactly one solution...the least squares solution!

Example 1 (revisited):

Find the least squares solution to the system of linear equations

$$\begin{bmatrix} -1 & 2 \\ 2 & -3 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 1 \\ 2 \end{bmatrix}.$$

Example 2

Find the least squares solution to the system of linear equations

$$\begin{bmatrix} 1 & -2 \\ -1 & 2 \\ 0 & 3 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ -4 \\ 2 \end{bmatrix}.$$

Example 3

Let us imagine that we are studying a physical system that gets hotter over time. Let us also suppose that we expect a linear relationship between time and temperature. That is, we expect time and temperature to be related by a formula of the form

$$T = at + b,$$

where T is temperature (in degrees Celsius), t is time (in seconds), and a and b are unknown physical constants. We wish to do an experiment to determine the (approximate) values for the constants a and b. We allow our system to get hot and measure the temperature at various times t. The following table summarizes our findings

t (sec)	0.5	1.1	1.5	2.1	2.3
T (°C)	32.0	33.0	34.2	35.1	35.7

.

Find the least squares solution to the linear system that arises from this experiment

$$\begin{cases} 0.5\,a + b &= 32.0 \\ 1.1\,a + b &= 33.0 \\ 1.5\,a + b &= 34.2 \\ 2.1\,a + b &= 35.1 \\ 2.3\,a + b &= 35.7 \end{cases} \quad \leftrightsquigarrow \quad \begin{bmatrix} 0.5 & 1 \\ 1.1 & 1 \\ 1.5 & 1 \\ 2.1 & 1 \\ 2.3 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 32.0 \\ 33.0 \\ 34.2 \\ 35.1 \\ 35.7 \end{bmatrix}.$$

Example 4

The table below is the estimated population of the United States (in millions) rounded to three digits. Suppose there is a linear relationship between time t and population $P(t)$. Use this data to predict the U.S. population in 2010.

year	1980	1985	1990	1995
population	227	237	249	262

.

Example 5 (Exponential Fit) (Example 4 Revisited)

In population studies, exponential models are much more commonly used than linear models. This means that we hope to find constants a and b such that the population $P(t)$ is given approximately by the equation $P(t) = a\,e^{bt}$. To convert this into a linear equation, we take the natural logarithm of both sides, producing

$$\ln P(t) = \ln a + bt.$$

Use the method of least squares to find values for $\ln a$ and b that best fit the data of Example 4.

year	1980	1985	1990	1995
ln(population)	5.425	5.468	5.517	5.568

.