

---

# Modeling with Data

Least Square Method  
(linear or nonlinear)

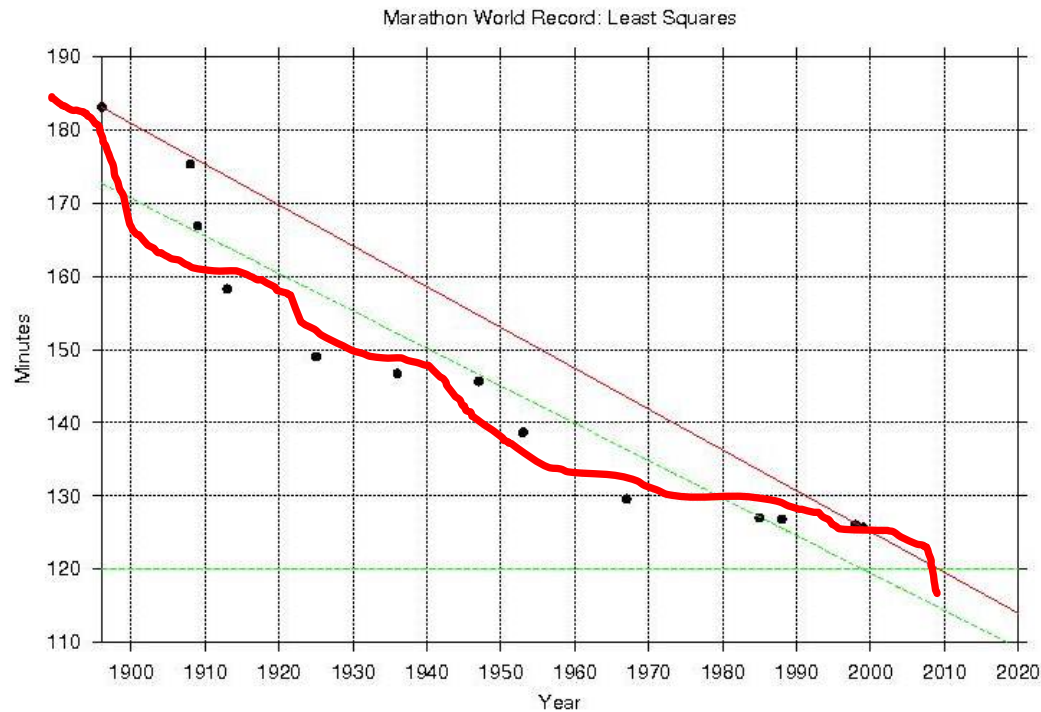
# Curve fitting

---

- ◆ Raw data usually has noise. The values of dependent variables vary even though all the independent variables are constant.
- ◆ Therefore, the estimation of the trend (the dependent variables) is needed. This process is called **regression** or **curve fitting**. The estimated equation (matrix) satisfy the raw data.
- ◆ However, the equation is **not usually unique**, and the equation or curve with a minimal deviation from all data points is desirable.
- ◆ This desirable best-fitting equation can be obtained by least square approximation method which uses the minimal sum of the deviations squared from a given set of data.

# Non Uniqueness of Fitting Curves

---



There are tons of ways to approximate given data, here we will focus on least square approximation.

# Least Square Approximation Method

---

If you have a data set  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  and the (trial) fitting curve (with unknown parameters)  $f(x)$  has the deviation  $d_1, d_2, \dots, d_n$  which are caused from each data point, the least square method is to determine the curve  $f(x)$  so that  $E$  has the minimum value;

$$E = \sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - f(x_i))^2$$

# Necessary condition

---

## Extreme Value Theory - Decision Theory

- If  $f(x_1, x_2, \dots, x_n)$  has an extreme value at  $(a_1, a_2, \dots, a_n)$ , then for  $i = 1, \dots, n$

$$\frac{\partial f}{\partial x_i}(a_1, a_2, \dots, a_n) = 0$$

# Examples of fitting curves

---

1. Linear function:  $ax + b$
  2. Quadratic function:  $ax^2 + bx + c$
  3. Cubic function:  $ax^3 + bx^2 + cx + d$
  4. Exponential function:  $ab^x$  or  $cx^\alpha$
- *We are going to determine the coefficients of the fitting curve which makes the least square error minimal.*
- \* In fact, there is no restrictions for a fitting curve.

# Linear function

---

- Data  $(x_i, y_i)$ ,  $i = 1, \dots, n$

- $f(x) = ax + b$

- $E(a, b) = \sum_{i=1}^n (f(x_i) - y_i)^2 = \sum_{i=1}^n (ax_i + b - y_i)^2$

- From  $\frac{\partial E}{\partial a} = 0$  and  $\frac{\partial E}{\partial b} = 0$ , we have

$$\frac{\partial E}{\partial a} = 2 \sum_{i=1}^n (ax_i + b - y_i)x_i = 0 \text{ and } \frac{\partial E}{\partial b} = 2 \sum_{i=1}^n (ax_i + b - y_i)$$

- Solving the above system of equations for  $a$  and  $b$ , we have

$$a = \frac{nS_{xy} - S_x S_y}{nS_{xx} - S_x^2}, \quad b = \frac{S_{xx} S_y - S_{xy} S_x}{nS_{xx} - S_x^2}$$

where  $S_{xx} = \sum_{i=1}^n x_i^2$ ,  $S_x = \sum_{i=1}^n x_i$ ,  $S_{xy} = \sum_{i=1}^n x_i y_i$ ,  $S_y = \sum_{i=1}^n y_i$ .

# Non Linear functions

Occasionally it is appropriate to assume that the data are exponentially related. This requires the approximating function to be of the form

$$y = be^{ax} \quad (8.4)$$

or

$$y = bx^a, \quad (8.5)$$

for some constants  $a$  and  $b$ . The difficulty with applying the least squares procedure in a situation of this type comes from attempting to minimize

$$E = \sum_{i=1}^m (y_i - be^{ax_i})^2, \quad \text{in the case of Eq. (8.4),}$$

or

$$E = \sum_{i=1}^m (y_i - bx_i^a)^2, \quad \text{in the case of Eq. (8.5).}$$

The normal equations associated with these procedures are obtained from either

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

and

$$0 = \frac{\partial E}{\partial a} = 2 \sum_{i=1}^m (y_i - be^{ax_i})(-bx_i e^{ax_i}), \quad \text{in the case of Eq. (8.4);}$$

or

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

and

$$0 = \frac{\partial E}{\partial a} = 2 \sum_{i=1}^m (y_i - bx_i^a)(-b(\ln x_i)x_i^a), \quad \text{in the case of Eq. (8.5).}$$

No exact solution to either of these systems in  $a$  and  $b$  can generally be found.

## An Idea

The method that is commonly used when the data are suspected to be exponentially related is to consider the logarithm of the approximating equation:

$$\ln y = \ln b + ax, \quad \text{in the case of Eq. (8.4),}$$

and

$$\ln y = \ln b + a \ln x, \quad \text{in the case of Eq. (8.5).}$$

In either case, a linear problem now appears, and solutions for  $\ln b$  and  $a$  can be obtained by appropriately modifying the normal equations (8.1) and (8.2).

However, the approximation obtained in this manner is *not* the least squares approximation for the original problem, and this approximation can in some cases differ significantly from the least squares approximation to the original problem. The application in

---

# Functions for Linear Least square

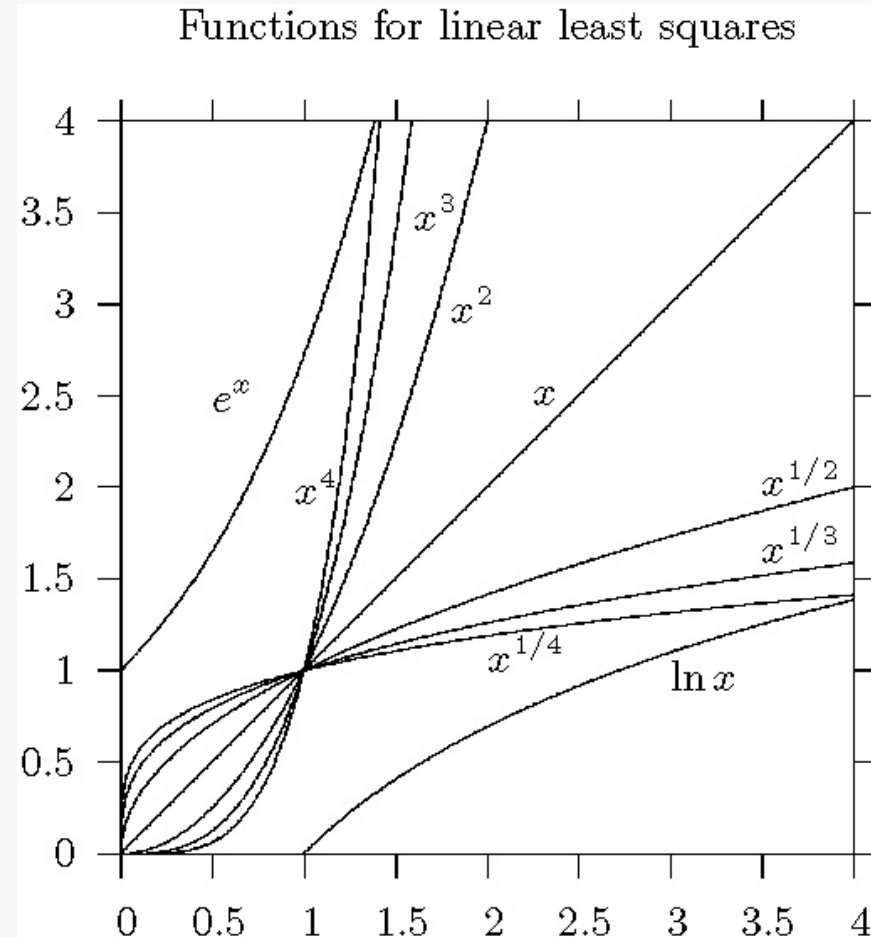
$$y = a + bx$$

$$y = ab^x \Rightarrow \log(y) = a + x \log(b)$$

$$y = cx^\alpha \Rightarrow \log(y) = \log(c) + \alpha \log(x)$$

$$y = ae^{bx} \Rightarrow \log(y) = \log(a) + bx$$

$\vdots$



# Some comments

---

- $\bar{y}$  = average of  $y_i$ ,  $i = 1, \dots, n$
- $SSR = \sum_{i=1}^n (f(x_i) - \bar{y})^2$ ,  $SST = \sum_{i=1}^n (y_i - \bar{y})^2$
- 결정계수 :  $R^2 = \frac{SSR}{SST}$ . When  $f(x)$  is linear,

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (f(x_i) - \bar{y})^2 + \sum_{i=1}^n (y_i - f(x_i))^2 \quad \dots \quad (1)$$

$$SST = SSR + SSE$$

종속변수의 전체 제곱 변동 ( $SST$ ) 중에 독립 변수 ( $SSR$ )  
에 의해 설명되는 비율을 의미

# Projects

---

- ◆ Express  $a, b, c$  in  $f(x)=ax^2+bx+c$  in terms of  $x_i$  and  $y_i$  ( $i=1,\dots,n$ ) as in the previous slide for linear least square method.
- ◆ Use linear least square method for trial functions  $ab^x$  and  $cx^\alpha$  to find formulae for  $a, b$  and  $c, \alpha$ .
- ◆ Show that the equation (1) holds in the previous slide.