Chapter 3:
Curve Fitting & Interpolation

Lecture 1:
Least-Squares Regression

Facebook's ad revenue:

2020: $84.5 billion*
2019: $69.4 billion*
2018: $55.0 billion
2017: $39.9 billion
2016: $26.9 billion
2015: $17.1 billion
2014: $11.5 billion
2013: $6.9 billion
2012: $4.3 billion
2011: $3.2 billion
2010: $1.9 billion
2009: $764 million

Source: Jon Erlichman Twitter

Motivation

Question: Estimate the Giant's revenue next year

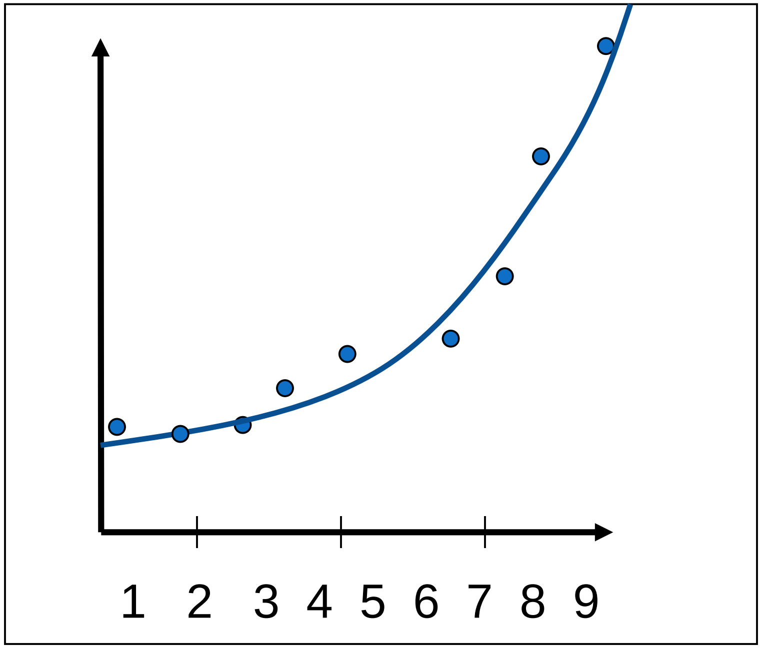
https://thanhnien.vn/cong-nghe/facebook-kiem-tien-nhu-the-nao-1174596.html

Motivation

- Data often given for discrete values
- Estimates at points between the discrete values

Experimental data:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$



The relationship between x and y is not known

Find a function $y=f(x)$ that best fits the data. Use this $f(x)$ to forecast about other values

Curve Fitting

Given a set of tabulated data, find a curve or a function that *best represents the data*.

Given:

1. The tabulated **data**
2. The **form** of the function
3. The curve fitting **criteria**

- Find the **unknown** coefficients of the function

Selection of the Functions

Linear $\quad f(x) = a + bx$

Quadratic $\quad f(x) = a + bx + cx^2$

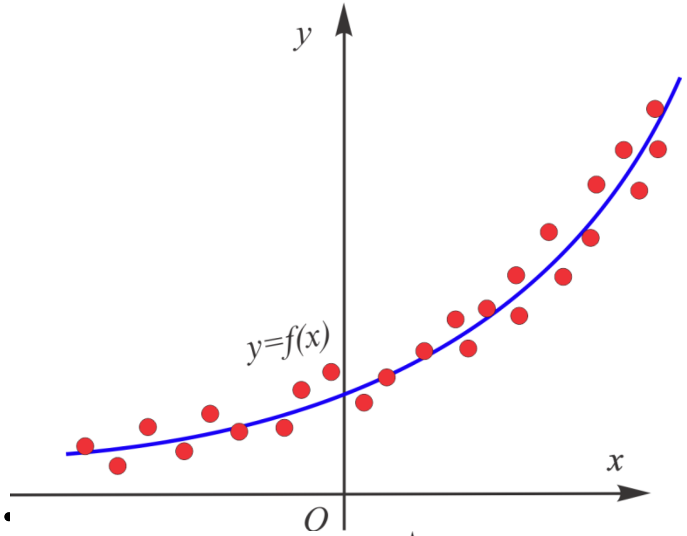
Polynomial $\quad f(x) = \sum_{k=0}^{n} a_k \, x^k$

General $\quad f(x) = \sum_{k=0}^{m} a_k \, g_k(x)$

$g_k(x)$ are known.

Decide on the Criterion

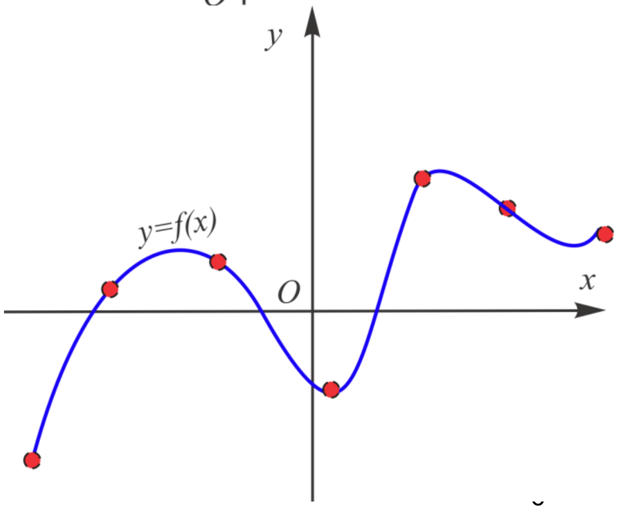
1. Least Squares Regression:



$$\text{minimize} \quad \Phi = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

2. Exact Matching (Interpolation):

$$y_i = f(x_i)$$



Least Squares Regression

Linear Regression

- Fitting a straight line to a set of paired observations: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

  $y = a + bx + e$

  $\boxed{Linear \qquad f(x) = a + bx}$

  a-intercept

  b-slope

  e-error, or residual, between the model and the observations.

Determine the Unknowns

We want to find a and b to minimize:

$$\Phi(a,b) = \sum_{i=1}^{n} (a + bx_i - y_i)^2$$

How do we obtain a and b to minimize: $\Phi(a,b)$?

Determine the Unknowns

Necessary condition for the minimum:

$$\boxed{\begin{gathered}\frac{\partial\, \Phi(a,b)}{\partial\, a} = 0 \\[1em] \frac{\partial\, \Phi(a,b)}{\partial\, b} = 0\end{gathered}}$$

Determining the Unknowns

$$\frac{\partial\ \Phi(a,b)}{\partial\ a} = \sum_{i=1}^{n} 2\left(a + bx_i - y_i\right) = 0$$

$$\frac{\partial\ \Phi(a,b)}{\partial\ b} = \sum_{i=1}^{n} 2\left(a + bx_i - y_i\right)x_i = 0$$

Normal Equations

$$\frac{\partial\, \Phi(a,b)}{\partial\, a} = \sum_{i=1}^{n} 2\left(a + bx_i - y_i\right) = 0$$

$$\frac{\partial\, \Phi(a,b)}{\partial\, b} = \sum_{i=1}^{n} 2\left(a + bx_i - y_i\right)x_i = 0$$

$$n\,a + \left(\sum_{i=1}^{n} x_i\right) b = \left(\sum_{i=1}^{n} y_i\right)$$

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

These equations are called Normal Equations

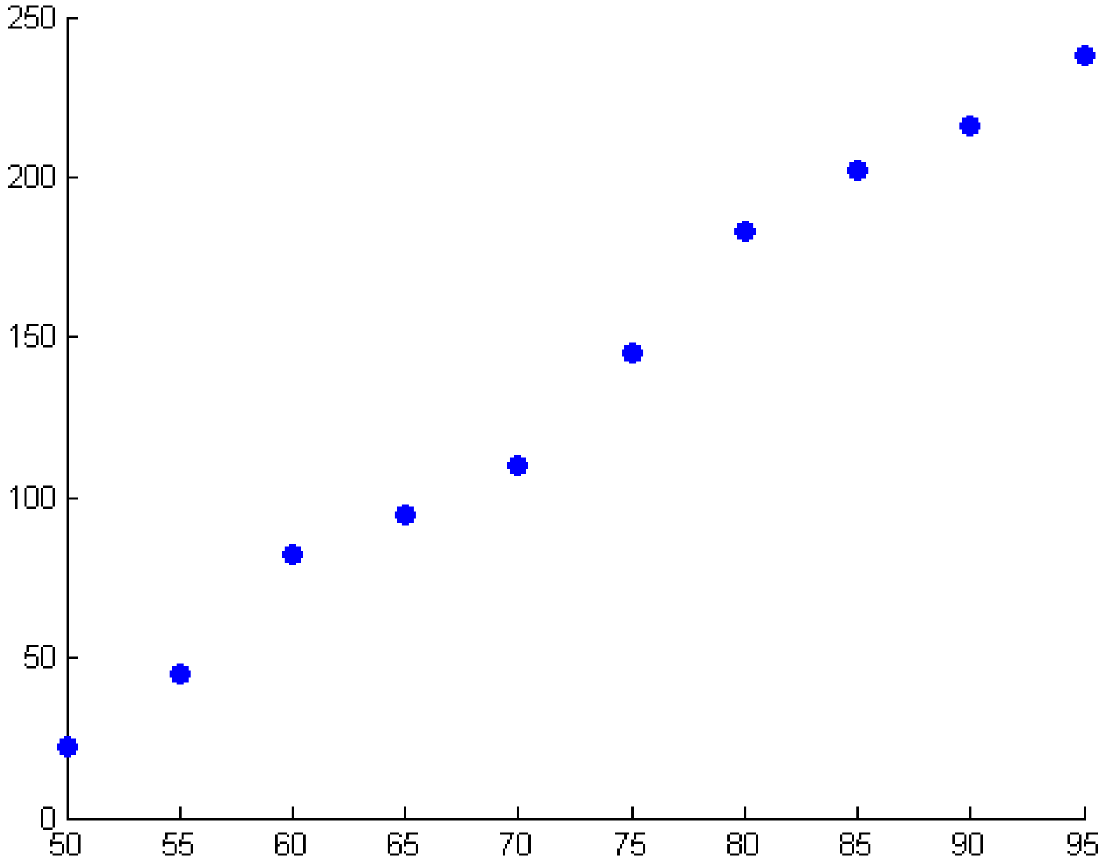
Example 1

- According to a research by a group of biologists, the chirping rate y (chirps/min) of crickets of a certain species appears to be related to temperature x (^{o}F).

x	50	55	60	65	70	75	80	85	90	95
y	22	45	82	95	110	145	183	202	216	238

- **Question:** What is the chirping rate at 100^{o}F?

Plot data



Solution

$$n\,a + \left(\sum_{k=1}^{n} x_k\right) b = \left(\sum_{k=1}^{n} y_k\right)$$

$$\left(\sum_{k=1}^{n} x_k\right) a + \left(\sum_{k=1}^{n} x_k^2\right) b = \left(\sum_{k=1}^{n} x_k\, y_k\right)$$

- Normal equations:

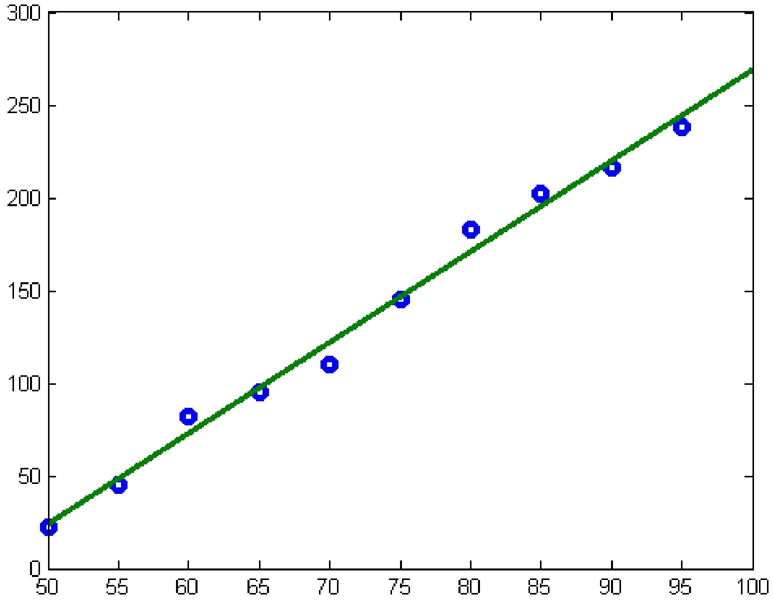
$$10a + 725b = 1338$$

$$725a + 54625b = 107105$$

$$\Rightarrow a = \text{-}221.2, \quad b = 4.897$$

$$y = \text{-}221.2 + 4.897x$$

$$y(100) = 268$$



Exercise

The profits y of a company (in millions USD) during the past seven years x are given in the following table

x	1	2	3	4	5	6	7
y	2.2	2.5	3.0	3.4	3.8	4.3	4.8

a) Plot the data points

b) Find the least-squares line that best fit the data. Plot the line with the data points

c) Use the least-squares line found in part b) to estimate the profit in the 8th year

$$na+\left(\sum_{k=1}^{n} x_k\right)b=\left(\sum_{k=1}^{n} y_k\right)$$

$$\left(\sum_{k=1}^{n} x_k\right)a+\left(\sum_{k=1}^{n} x_k^2\right)b=\left(\sum_{k=1}^{n} x_k\, y_k\right)$$

Quiz3

The profits y of a company (in millions USD) during the past seven years x are given in the following table

x	1	2	3	4	5	6	7
y	1.2	1.5	2.0	2.5	3.6	4.6	6.8

a) Develop a least-squares, a quadratic models that best fit the data. Plot these functions with the data points

b) Use the least-squares models found in part a) to estimate the profit in the 8^{th} year

c) Which model in a) is the best one? Why?

Multiple Linear Regression

Example:

Given the following data:

t	0	1	2	3
x	0.1	0.4	0.2	0.2
y	3	2	1	2

Determine a function of two variables:

$$f(x,t) = a + b\,x + c\,t$$

That best fits the data with the least sum of the square of errors.

Solution of Multiple Linear Regression

Construct $\Phi = \sum_{i=1}^{n} (y_i - f(x_i, t_i))^2$

t	0	1	2	3
x	0.1	0.4	0.2	0.2
y	3	2	1	2

Derive the necessary conditions by equating the partial derivatives with respect to the unknown parameters to zero, then solve the equations.

Solution of Multiple Linear Regression

$$f(x,t) = a + bx + ct, \quad \Phi(a,b,c) = \sum_{i=1}^{n} \left(a + bx_i + ct_i - y_i\right)^2 \to \min$$

Necessary conditions:

$$\frac{\partial \Phi(a,b,c)}{\partial a} = 2\sum_{i=1}^{n} \left(a + bx_i + ct_i - y_i\right) = 0$$

$$\frac{\partial \Phi(a,b,c)}{\partial b} = 2\sum_{i=1}^{n} \left(a + bx_i + ct_i - y_i\right) x_i = 0$$

$$\frac{\partial \Phi(a,b,c)}{\partial c} = 2\sum_{i=1}^{n} \left(a + bx_i + ct_i - y_i\right) t_i = 0$$

Normal Equations

$$a\,n + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} (x_i)^2 + c\sum_{i=1}^{n} (x_i\; t_i) = \sum_{i=1}^{n} (x_i\; y_i)$$

$$a\sum_{i=1}^{n} t_i + b\sum_{i=1}^{n} (x_i\; t_i) + c\sum_{i=1}^{n} (t_i)^2 = \sum_{i=1}^{n} (t_i\; y_i)$$

Example 2: Multiple Linear Regression

i	1	2	3	4	Sum
t_i	0	1	2	3	6
x_i	0.1	0.4	0.2	0.2	0.9
y_i	3	2	1	2	8
x_i^2	0.01	0.16	0.04	0.04	0.25
$x_i\ t_i$	0	0.4	0.4	0.6	1.4
$x_i\ y_i$	0.3	0.8	0.2	0.4	1.7
t_i^2	0	1	4	9	14
$t_i\ y_i$	0	2	2	6	10

Example 2: System of Equations

$$4a + 0.9b + 6c = 8$$

$$0.9a + 0.25b + 1.4c = 1.7$$

$$6a + 1.4b + 14c = 10$$

$$Solving:$$

$$a = 2.9574, \; b = -1.7021, \;\; c = -0.38298$$

$$f(x,t) = a + bx + ct = 2.9574 - 1.7021\,x - 0.38298\,t$$

Exercise

Develop a multiple least-square model for these data, and use it to estimate y(1/2, -1)

t	-1	1	3	4
x	-2	0	1	3
y	-6	-2	3	6

Normal Equations:

$$a\,n + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} (x_i)^2 + c\sum_{i=1}^{n} (x_i\ t_i) = \sum_{i=1}^{n} (x_i\ y_i)$$

$$a\sum_{i=1}^{n} t_i + b\sum_{i=1}^{n} (x_i\ t_i) + c\sum_{i=1}^{n} (t_i)^2 = \sum_{i=1}^{n} (t_i\ y_i)$$

$$x = (x_1, x_2, \ldots, x_n), t = (t_1, t_2, \ldots, t_n)$$

$$y = (y_1, y_2, \ldots, y_n)$$

$$a\ n + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + bx^2 + cxt = xy$$

$$a\sum_{i=1}^{n} t_i + bxt + ct^2 = ty$$

Exercise 2

Develop a multiple least-square model for these data, and use it to estimate y(x=3, t=5)

t	0	1	3	4
x	1	2	3	4
y	2	3	5	8

Normal Equations:

$$a\,n + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n}(x_i)^2 + c\sum_{i=1}^{n}(x_i\ t_i) = \sum_{i=1}^{n}(x_i\ y_i)$$

$$a\sum_{i=1}^{n} t_i + b\sum_{i=1}^{n}(x_i\ t_i) + c\sum_{i=1}^{n}(t_i)^2 = \sum_{i=1}^{n}(t_i\ y_i)$$

$$x = (x_1, x_2, \ldots, x_n), t = (t_1, t_2, \ldots, t_n)$$

$$y = (y_1, y_2, \ldots, y_n)$$

$$a\ n + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + bx^2 + cxt = xy$$

$$a\sum_{i=1}^{n} t_i + bxt + ct^2 = ty$$

Nonlinear Least Squares Models and Methods

- Polynomial Regression
- Fitting with Nonlinear Functions
- Linearization Method

Polynomial Regression

- The least squares method can be extended to fit the data to a higher-order polynomial

$$f(x) = a + bx + cx^2, \quad e_i^2 = (f(x_i) - y_i)^2$$

$$\text{Minimize} \quad \Phi(a,b,c) = \sum_{i=1}^{n} \left(a + bx_i + cx_i^2 - y_i \right)^2$$

Necessary conditions:

$$\frac{\partial \Phi(a,b,c)}{\partial a} = 0, \qquad \frac{\partial \Phi(a,b,c)}{\partial b} = 0, \qquad \frac{\partial \Phi(a,b,c)}{\partial c} = 0$$

Equations for Quadratic Regression

$$\text{Minimize} \quad \Phi(a,b,c) = \sum_{i=1}^{n} \left(a + bx_i + cx_i^2 - y_i \right)^2$$

$$\frac{\partial \Phi(a,b,c)}{\partial a} = 2\sum_{i=1}^{n} \left(a + bx_i + cx_i^2 - y_i \right) = 0$$

$$\frac{\partial \Phi(a,b,c)}{\partial b} = 2\sum_{i=1}^{n} \left(a + bx_i + cx_i^2 - y_i \right) x_i = 0$$

$$\frac{\partial \Phi(a,b,c)}{\partial c} = 2\sum_{i=1}^{n} \left(a + bx_i + cx_i^2 - y_i \right) x_i^2 = 0$$

Normal Equations

$$a\,n + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i y_i$$

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2 y_i$$

Example 3: Polynomial Regression

Fit a second-order polynomial to the following data

x_i	0	1	2	3	4	5	$\sum=15$
y_i	2.1	7.7	13.6	27.2	40.9	61.1	$\sum=152.6$
x_i^2	0	1	4	9	16	25	$\sum=55$
x_i^3	0	1	8	27	64	125	225
x_i^4	0	1	16	81	256	625	$\sum=979$
$x_i\ y_i$	0	7.7	27.2	81.6	163.6	305.5	$\sum=585.6$
$x_i^2\ y_i$	0	7.7	54.4	244.8	654.4	1527.5	$\sum=2488.8$

Example 3: Equations and Solution

$$6\,a + 15\,b + 55\,c = 152.6$$

$$15\,a + 55\,b + 225\,c = 585.6$$

$$55\,a + 225\,b + 979\,c = 2488.8$$

Solving…

$$a = 2.4786,\ \ b = 2.3593,\ \ c = 1.8607$$

$$f(x) = 2.4786 + 2.3593\,x + 1.8607\,x^2$$

Justify the best choice

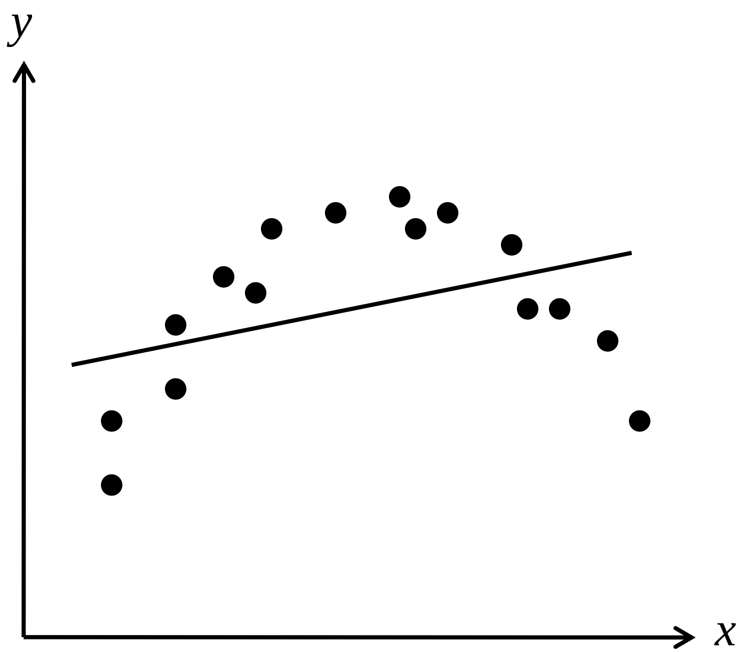
Given two or more functions to fit the data,
How do you select the best?

Answer :

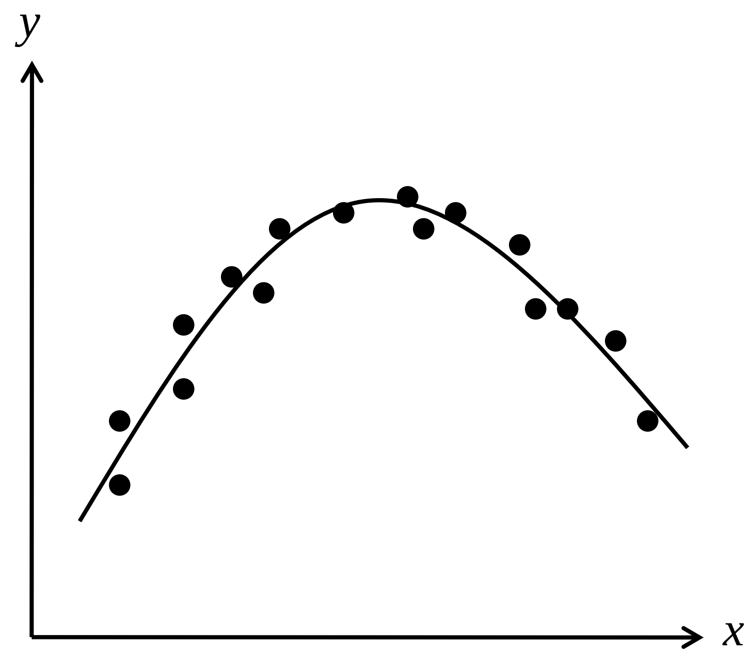
Determine the parameters for each function, then compute Φ for each one. The function resulting in smaller Φ (least sum of the squares of the errors) is the best.

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

Example showing that Quadratic is preferable than Linear Regression



Linear Regression



Quadratic Regression

Fitting with Nonlinear Functions

x_i	0.24	0.65	0.95	1.24	1.73	2.01	2.23	2.52
y_i	0.23	-0.23	-1.1	-0.45	0.27	0.1	-0.29	0.24

It is required to find a function of the form, for example:

$$f(x) = a\ \ln(x) + b\ \cos(x) + c\ e^{x}$$

to fit the data.

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

General $\quad f(x) = \sum_{k=0}^{m} a_k\ g_k(x)$

$g_k(x)$ *are known.*

Fitting with Nonlinear Functions

$$\Phi(a,b,c) = \sum_{i=1}^{n}(a\ln(x_i) + b\cos(x_i) + c\,e^{x_i} - y_i)^2$$

Necessary condition for the minimum:

$$\left.\begin{array}{l} \dfrac{\partial\,\Phi(a,b,c)}{\partial\,a} = 0 \\[2ex] \dfrac{\partial\,\Phi(a,b,c)}{\partial\,b} = 0 \\[2ex] \dfrac{\partial\,\Phi(a,b,c)}{\partial\,c} = 0 \end{array}\right\} \Rightarrow \textit{Normal Equations}$$

Normal Equations

$$a\sum_{i=1}^{n}(\ln x_i)^2 + b\sum_{i=1}^{n}(\ln x_i)(\cos x_i) + c\sum_{i=1}^{n}(\ln x_i)(e^{x_i}) = \sum_{i=1}^{n} y_i(\ln x_i)$$

$$a\sum_{i=1}^{n}(\ln x_i)(\cos x_i) + b\sum_{i=1}^{n}(\cos x_i)^2 + c\sum_{i=1}^{n}(\cos x_i)(e^{x_i}) = \sum_{i=1}^{n} y_i(\cos x_i)$$

$$a\sum_{i=1}^{n}(\ln x_i)(e^{x_i}) + b\sum_{i=1}^{n}(\cos x_i)(e^{x_i}) + c\sum_{i=1}^{n}(e^{x_i})^2 = \sum_{i=1}^{n} y_i(e^{x_i})$$

Evaluate the sums and solve the normal equations.

Example 4: Evaluating Sums

xi	0.24	0.65	0.95	1.24	1.73	2.01	2.23	2.52	∑=11.57
yi	0.23	-0.23	-1.1	-0.45	0.27	0.1	-0.29	0.24	∑=-1.23
$(\ln xi)^2$	2.036	0.1856	0.0026	0.0463	0.3004	0.4874	0.6432	0.8543	∑=4.556
ln(xi) cos(xi)	-1.386	-0.3429	-0.0298	0.0699	-0.0869	-0.2969	-0.4912	-0.7514	∑=-3.316
$\ln(xi) * e^{xi}$	-1.814	-0.8252	-0.1326	0.7433	3.0918	5.2104	7.4585	11.487	∑=25.219
yi * ln(xi)	-0.328	0.0991	0.0564	-0.0968	0.1480	0.0698	-0.2326	0.2218	∑=-0.0625
$\cos(xi)^2$	0.943	0.6337	0.3384	0.1055	0.0251	0.1808	0.3751	0.6609	∑=3.26307
$\cos(xi) * e^{xi}$	1.235	1.5249	1.5041	1.1224	-0.8942	-3.1735	-5.696	-10.104	∑=-14.481
yi*cos(xi)	0.223	-0.1831	-0.6399	-0.1462	-0.0428	-0.0425	0.1776	-0.1951	∑=-0.8485
$(e^{xi})^2$	1.616	3.6693	6.6859	11.941	31.817	55.701	86.488	154.47	∑=352.39
$yi * e^{xi}$	0.2924	-0.4406	-2.844	-1.555	1.523	0.7463	-2.697	2.9829	∑=-1.9923

Example 4: Equations & Solution

$$4.55643\,a - 3.31547\,b + 25.2192\,c = -0.062486$$

$$-3.31547\,a + 3.26307\,b - 14.4815\,c = -0.848514$$

$$25.2192\,a - 14.4815\,b + 352.388\,c = -1.992283$$

Solving the above equations :

$$a = -0.88815, \quad b = -1.1074, \quad c = 0.012398$$

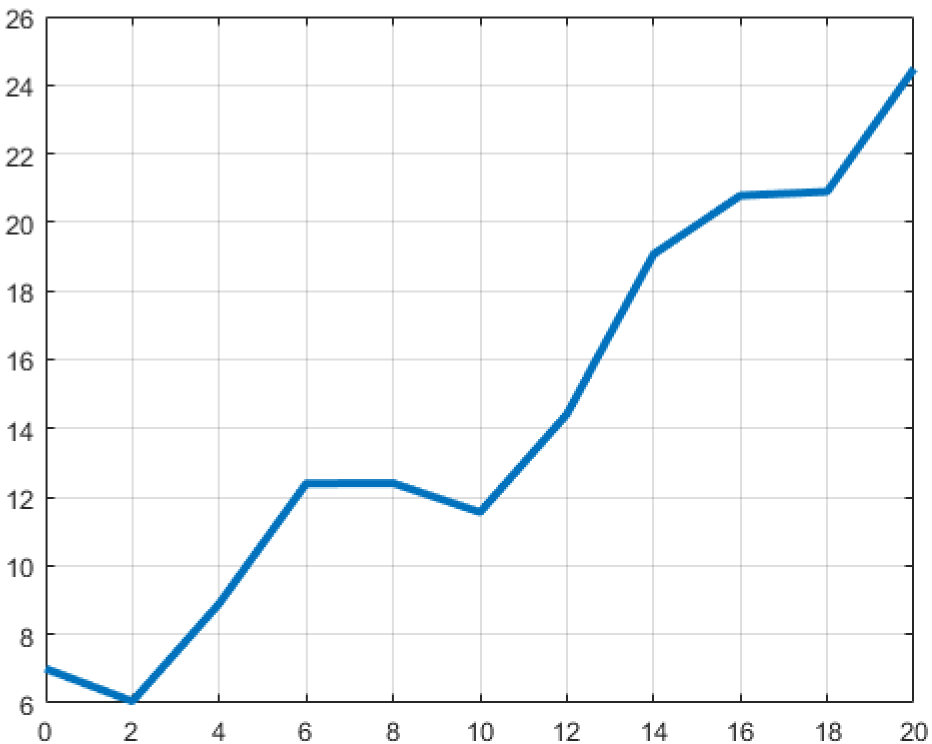
Therefore,

$$f(x) = -0.88815\ \ln(x) - 1.1074\cos(x) + 0.012398\,e^{x}$$

Exercise: Price of a product in the last 9 years as:

7.0 6.05 8.9 12.4 12.4 11.6 14.4 19.1 20.8 20.9

$$y = 2e^{0.1t} - 2\sin(0.75t) + 0.3t + 5$$



Using least-squares regression with the following nonlinear model to predict the price for the next year

$$y = ae^{0.1t} + b\sin(0.75t) + ct + 5$$

Linearization Method: Exponential Model

Example: Given data

x_i	1	2	3
y_i	2.4	5	9

Find a function $f(x) = ae^{bx}$ that best fits the data.

$$\Phi(a,b) = \sum_{i=1}^{n} \left(ae^{bx_i} - y_i\right)^2$$

Normal Equations are obtained using:

$$\frac{\partial \Phi}{\partial a} = 2\sum_{i=1}^{n} \left(ae^{bx_i} - y_i\right)e^{bx_i} = 0$$

Difficult to Solve

$$\frac{\partial \Phi}{\partial b} = 2\sum_{i=1}^{n} \left(ae^{bx_i} - y_i\right)a\, x_i\, e^{bx_i} = 0$$

Exponential Model

Find a function $f(x) = ae^{bx}$ that best fits the data.

Then $\ln(f(x)) = \ln(a) + b\,x$

Define $z = \alpha + bx$

where $\alpha = \ln(a)$ and $z = \ln(y)$

Instead of minimizing: $\Phi(a,b) = \sum_{i=1}^{n} \left(ae^{bx_i} - y_i \right)^2$

Minimize: $\Phi(\alpha,b) = \sum_{i=1}^{n} \left(\alpha + bx_i - z_i \right)^2$ (Easier to solve)

Normal Equations

$$\Phi(\alpha,b)=\sum_{i=1}^{n}\left(\alpha+b\,x_i-z_i\right)^2$$

Normal Equations are obtained using:

$$\frac{\partial\Phi}{\partial\alpha}=2\sum_{i=1}^{n}\left(\alpha+b\,x_i-z_i\right)=0$$

$$\frac{\partial\Phi}{\partial b}=2\sum_{i=1}^{n}\left(\alpha+b\,x_i-z_i\right)x_i=0$$

$$\alpha\,n+b\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}z_i \quad \text{and} \quad \alpha\sum_{i=1}^{n}x_i+b\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}(x_i\ z_i)$$

Evaluating Sums and Solving

x_i	1	2	3	Σ=6
y_i	2.4	5	9	
$z_i=\ln(y_i)$	0.875469	1.609438	2.197225	Σ=4.68213
x_i^2	1	4	9	Σ=14
$x_i\ z_i$	0.875469	3.218876	6.591674	Σ=10.6860

Set $z = \ln y$

Normal Equations for x and z:

$$3\ \alpha + 6\ b = 4.68213$$

$$6\ \alpha + 14\ b = 10.686$$

Solving Equations:

$$\alpha = 0.23897, \quad b = 0.66087$$

$$\alpha = \ln(a), \quad a = e^{\alpha}$$

$$a = e^{0.23897} = 1.26994$$

$$f(x) = ae^{bx} = 1.26994\, e^{0.66087x}$$

Linearization Method: Other Models

1. Power equation $y = ax^b$

Take logarithm $\log y = \log a + b \log x$

2. Saturation-growth-rate equation $y = \frac{ax}{b+x}$

Inverting it: $\frac{1}{y} = \frac{b}{a}\frac{1}{x} + \frac{1}{a}$

Homework N3. **Deadline**: 5 weeks

Problem 1: The profit (in millions USD) of a company during the past seven years are reported in the following table

Time (years)	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Profit (millions $)	1.2	1.5	1.8	2.m	3.8	4.7	6.n

- Find a best-fit equation to the data trend. Try several possibilities- linear, parabolic, and exponential. Plot all these curves and the data in the same coordinates. Find the best equation to predict the profit in two years.
-

(m-2)(n-2) is the two last digits of your student ID number

Homework N3: Problem 2

3. Use linear regression to fit these data by a plane

x	1	2	3	3+1/m	4	6
t	1	2	2+1/n	3	4	5
y	5	12	9	4	3	24

Use this plane to estimate f(2,3)

Homework N3: Problem 3

Given the data

x	**0**	**1**	**3**	**4**
y	1+1/n	2.2	4.5	10+1/m

a) Find the Newton divided-difference interpolating polynomial that passes through these data points. Use it to estimate f(2). Plot the curve.

b) Find the Lagrange interpolating polynomial that passes through these data points. Use it to estimate f'(2)

(m-2)(n-2) is the two last digits of your student ID number

Homework N3

Problem 4: Fit the data in the following Table with quadratic splines, and cubic splines. Use the result to evaluate the function at x=1.5

x	**0**	**1**	**2**	**3**
f(x)	1+1/n	2	4-1/m	12

Problem 5: Reconstruct the function $f(x) = [\frac{m+1+x^2}{n+1}]^x$ in [-3, 2] by linear splines, quadratic splines, and cubic splines using the values of *f(x)* at *x*=-3, -2, -1, 0, and 2. Use these splines to estimate the value *f*(1), and then find the errors

S. Chapra & R.P. Canale, Numerical Methods for Engineers, McGraw-Hill, 7th ed., 2015:
Pages 575-578
Problems: 20.20, 20.21, 20.25 20.26, 20.32, 20.33