MATH 3290 Mathematical Modeling

Chapter 4: Experimental Modeling

Kuang HUANG

February 1, 2024

Department of Mathematics
The Chinese University of Hong Kong

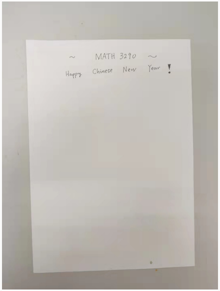
Course webpage

https://www.math.cuhk.edu.hk/course/2324/math3290

SCAN ME

About assignments

- The assignments will be posted on Blackboard next week.
- The first assignment is due by 5pm, Feb. 20.
- Writing everything with LaTeX if you major in science.
- You may also submit scanned assignments to Blackboard.

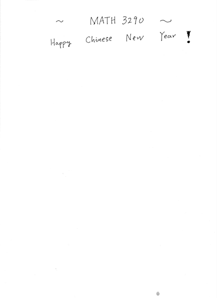


$\Rightarrow$

Use your phone with scanner apps such as

Simple Scan - PDF Scanner App

$\Rightarrow$



Introduction

We will construct empirical models based on the given data.

In Chap. 3, we construct a model by first assuming a particular type of functions, and then fit the model to the data.

Key assumption: we need to have some knowledge about what types of models are suitable.

In this chapter, we will construct empirical models:

- We do not assume that the model functions belong to a certain type.
- The model is determined solely by the data.

One-term models

Given a set of data points (x_i, y_i), our goal is to fit them to a model.

Q: how do we determine a suitable model function?

A: try 😅 😅 😅

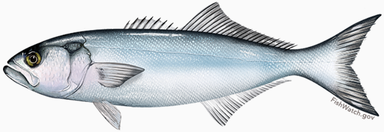
Main idea:

- Select functions $f(x)$ and $g(y)$ (e.g. the Tukey ladder of powers x^2, x, $\sqrt{x}$, $\ln(x)$, $1/\sqrt{x}$, $1/x$, $1/x^2$,...,);
- plot $g(y_i)$ vs $f(x_i)$;
- look for a linear relationship;
- use the model function $g(y) = af(x) + b$, determine a and b;
- if not, try other $f(x)$ and $g(y)$.

Example: bluefish population

Consider the data set.

Year	Bluefish (lb)
1940	15,000
1945	150,000
1950	250,000
1955	275,000
1960	270,000
1965	280,000
1970	290,000
1975	650,000
1980	1,200,000
1985	1,500,000
1990	2,750,000



Bluefish

Remark: we can change the unit of y from lb to 10^4 lb.

Example: bluefish population

Consider the data set.

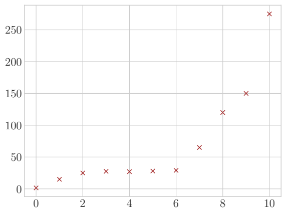
Year	Bluefish (lb)
1940	15,000
1945	150,000
1950	250,000
1955	275,000
1960	270,000
1965	280,000
1970	290,000
1975	650,000
1980	1,200,000
1985	1,500,000
1990	2,750,000

We take $f(x) = x$ and consider 4 cases:

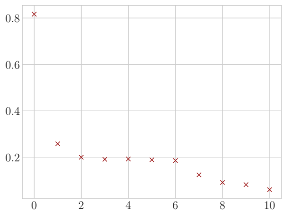
- $g(y) = y$,
- $g(y) = 1/\sqrt{y}$,
- $g(y) = \sqrt{y}$,
- $g(y) = \ln(y)$.

We plot $g(y_i)$ vs $f(x_i)$.

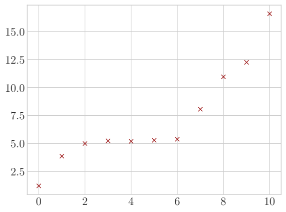
Remark: we can change the unit of y from lb to 10^4 lb.



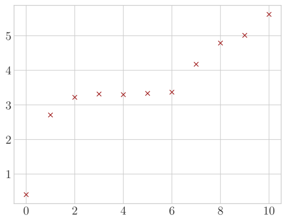
$g(y) = y$



$g(y) = 1/\sqrt{y}$



$g(y) = \sqrt{y}$



$g(y) = \ln(y)$

Hence, we will fit the model function

$$\sqrt{y} = ax + b$$

to the given data.

We let $\tilde{y} = \sqrt{y}$.

From Chap. 3, we need to solve

$$\begin{aligned} a\left(\sum_{i=1}^{m} x_i^2\right) + b\left(\sum_{i=1}^{m} x_i\right) &= \sum_{i=1}^{m} x_i \tilde{y}_i, \\ a\left(\sum_{i=1}^{m} x_i\right) + b\left(\sum_{i=1}^{m} 1\right) &= \sum_{i=1}^{m} \tilde{y}_i. \end{aligned}$$

Using the data set

$$\sum_{i=1}^{m} x_i^2 = 385, \quad \sum_{i=1}^{m} x_i = 55, \quad \sum_{i=1}^{m} 1 = 11,$$

$$\sum_{i=1}^{m} x_i \tilde{y}_i = 529.28, \quad \sum_{i=1}^{m} \tilde{y}_i = 79.06.$$

The linear system is

$$385a + 55b = 529.28, \quad 55a + 11b = 79.06.$$

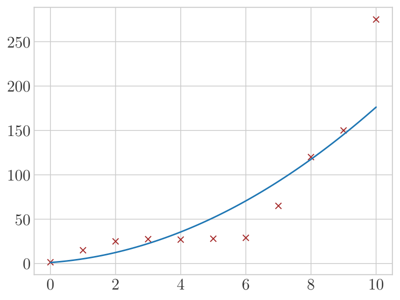
Solving it, we have $a = 1.21$ and $b = 1.09$.

The model is $\tilde{y} = 1.21x + 1.09$.

Therefore, we have $y = (1.21x + 1.09)^2$.

Year	Bluefish (lb)
1940	15,000
1945	150,000
1950	250,000
1955	275,000
1960	270,000
1965	280,000
1970	290,000
1975	650,000
1980	1,200,000
1985	1,500,000
1990	2,750,000

The given data set



The model function

For example, one can predict the bluefish population in 1995.

Let $x = 11$. Then $y = 210.11$. The bluefish population is 2,101,100lb.

Example: temperature distribution

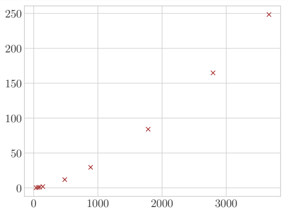
Assume you measure the temperature Y of a rod at various locations X, and obtain the following data.

Observation number	X	Y
1	35.97	0.241
2	67.21	0.615
3	92.96	1.000
4	141.70	1.881
5	483.70	11.860
6	886.70	29.460
7	1783.00	84.020
8	2794.00	164.800
9	3666.00	248.400

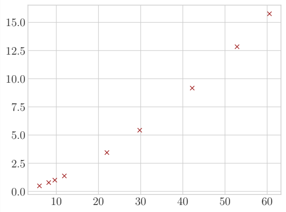
Consider 4 cases:

1. $f(x) = x$, $g(y) = y$;
2. $f(x) = \sqrt{x}$, $g(y) = \sqrt{y}$;
3. $f(x) = \ln(x)$, $g(y) = \sqrt{y}$;
4. $f(x) = \ln(x)$, $g(y) = \ln(y)$.

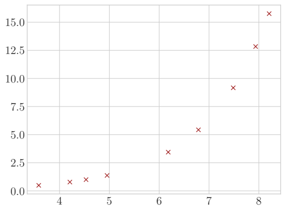
We plot $g(y_i)$ vs $f(x_i)$.



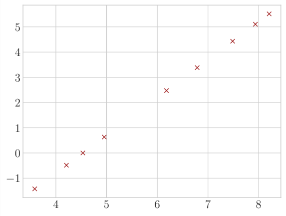
$f(x) = x, g(y) = y$



$f(x) = \sqrt{x}, g(y) = \sqrt{y}$



$f(x) = \ln(x), g(y) = \sqrt{y}$



$f(x) = \ln(x), g(y) = \ln(y)$

Hence, we will fit the model function

$$\ln(y) = a\ln(x) + b$$

to the given data.

We let $\tilde{x} = \ln(x)$ and $\tilde{y} = \ln(y)$.

From Chap. 3, we need to solve

$$\begin{aligned} a\Big(\sum_{i=1}^{m} \tilde{x}_i^2\Big) + b\Big(\sum_{i=1}^{m} \tilde{x}_i\Big) &= \sum_{i=1}^{m} \tilde{x}_i \tilde{y}_i, \\ a\Big(\sum_{i=1}^{m} \tilde{x}_i\Big) + b\Big(\sum_{i=1}^{m} 1\Big) &= \sum_{i=1}^{m} \tilde{y}_i. \end{aligned}$$

Using the data set

$$\sum_{i=1}^{m} \tilde{x}_i^2 = 346.26, \quad \sum_{i=1}^{m} \tilde{x}_i = 53.87, \quad \sum_{i=1}^{m} 1 = 9,$$

$$\sum_{i=1}^{m} \tilde{x}_i \tilde{y}_i = 153.18, \quad \sum_{i=1}^{m} \tilde{y}_i = 19.63.$$

The linear system is

$$346.26a + 53.87b = 153.18, \quad 53.87a + 9b = 19.63.$$

Solving it, we have $a = 1.500$ and $b = -6.798$.

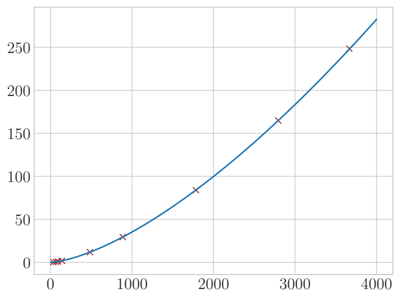
The model is $\tilde{y} = 1.500\tilde{x} - 6.798$.

Therefore, we have $\ln(y) = 1.500 \ln(x) - 6.798$.

That is $y = e^{-6.798} x^{1.500}$.

Observation number	X	Y
1	35.97	0.241
2	67.21	0.615
3	92.96	1.000
4	141.70	1.881
5	483.70	11.860
6	886.70	29.460
7	1783.00	84.020
8	2794.00	164.800
9	3666.00	248.400

The given data set

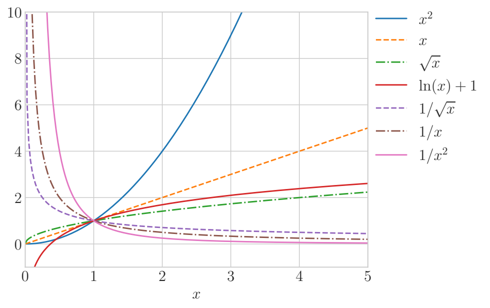


The model function

For example, one can predict the temperature at position $X = 3000$.

Let $x = 3000.00$. Then $y = 183.470$. Temperature $Y = 183.470$.

Facts about one-term models



The Tukey ladder of powers

- Note that functions in the Tukey ladder of powers are all increasing or decreasing.
- Then $y = g^{-1}(af(x) + b)$ is either increasing or decreasing.
- One-term models are not suitable for non-monotonic data patterns.

High-order polynomial models

A disadvantage of one-term models: too simple to capture complicated trend in the data.

In this part, we consider high-order polynomial models.

We obtain a function that goes through all data points.

Advantages of high-order polynomials: easy to differentiate and integrate.

E.g. one can find the maximum temperature (differentiation).

E.g. one can find the distance from the speed (integration).

Example: elapsed time of a tape recorder

We collected data relating the counter c on a tape recorder with its elapsed playing time t.

c_i	100	200	300	400	500	600	700	800
t_i (sec)	205	430	677	945	1233	1542	1872	2224

Example: elapsed time of a tape recorder

We collected data relating the counter c on a tape recorder with its elapsed playing time t.

c_i	100	200	300	400	500	600	700	800
t_i (sec)	205	430	677	945	1233	1542	1872	2224

We construct an empirical model using a high-order polynomial. Moreover, note that c is the independent variable.

We will find a 7-th order polynomial, denoted $P_7(c)$, passing through all data points.

$$P_7(c) = a_0 + a_1 c + a_2 c^2 + a_3 c^3 + a_4 c^4 + a_5 c^5 + a_6 c^6 + a_7 c^7$$

Recall, we have the data set:

c_i	100	200	300	400	500	600	700	800
t_i (sec)	205	430	677	945	1233	1542	1872	2224

We need that $P_7(c)$ goes through all data points:

$$
\begin{aligned}
205 &= a_0 + 1a_1 + 1^2a_2 + 1^3a_3 + 1^4a_4 + 1^5a_5 + 1^6a_6 + 1^7a_7 \\
430 &= a_0 + 2a_1 + 2^2a_2 + 2^3a_3 + 2^4a_4 + 2^5a_5 + 2^6a_6 + 2^7a_7 \\
&\vdots \\
2224 &= a_0 + 8a_1 + 8^2a_2 + 8^3a_3 + 8^4a_4 + 8^5a_5 + 8^6a_6 + 8^7a_7
\end{aligned}
$$

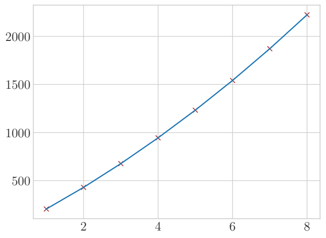
Note:

- We change the unit of c.
- We obtain a system of 8 linear equations.
- This is the so-called Vandermonde system.

Solving the above linear system:

$$\begin{aligned} a_0 &= -13.9999923 & a_4 &= -5.354166491 \\ a_1 &= 232.9119031 & a_5 &= 0.8013888621 \\ a_2 &= -29.08333188 & a_6 &= -0.0624999978 \\ a_3 &= 19.78472156 & a_7 &= 0.0019841269 \end{aligned}$$

The following plot is about $P_7(c)$ and the data.



Lagrangian form of polynomial

Given a set of $(n+1)$ data points (x_i, y_i), $i = 0, 1, 2, \ldots, n$, we need to find a polynomial $P(x)$ of degree n passing through all data points.

It is difficult to solve a large linear system of $(n+1) \times (n+1)$.

We can conveniently find $P(x)$ using Lagrangian bases:

$$L_k(x) = \prod_{i=0, i \neq k}^{n} \frac{x - x_i}{x_k - x_i}.$$

The $P(x)$ can be written as

$$P(x) = y_0 L_0(x) + y_1 L_1(x) + \cdots + y_n L_n(x).$$

Note:

$$L_k(x_k) = 1, \qquad L_k(x_j) = 0, \ j \neq k.$$

Example

Consider the data set (there are 4 data points):

x	x_1	x_2	x_3	x_4
y	y_1	y_2	y_3	y_4

We need to find a 3-rd order polynomial.

Using the above Lagrangian bases, we have

$$\begin{aligned} P_3(x) = {} & \frac{(x-x_2)(x-x_3)(x-x_4)}{(x_1-x_2)(x_1-x_3)(x_1-x_4)} y_1 + \frac{(x-x_1)(x-x_3)(x-x_4)}{(x_2-x_1)(x_2-x_3)(x_2-x_4)} y_2 \\ & + \frac{(x-x_1)(x-x_2)(x-x_4)}{(x_3-x_1)(x_3-x_2)(x_3-x_4)} y_3 + \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_4-x_1)(x_4-x_2)(x_4-x_3)} y_4 \end{aligned}$$

Advantages and disadvantages

Constructing an empirical model by a high-order polynomial—

Advantages:

- is "usually" easy to write down (using Lagrangian bases),
- has a better ability to capture complicated trends (cf. one-term models),
- can be differentiated and integrated easily.

However, it may—

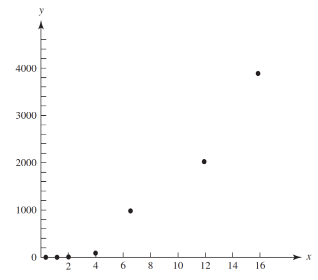
Disadvantages:

- contain too many oscillations (see Example 1),
- be very sensitive to errors in the data (see Example 2).

Example 1

Consider the following data set.

x	0.55	1.2	2	4	6.5	12	16
y	0.13	0.64	5.8	102	210	2030	3900



The data suggests that, the model function should be an increasing function.

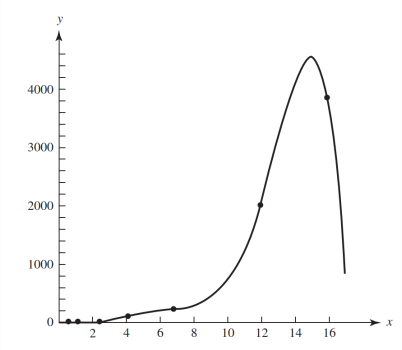
Assume that we construct a 6-th order polynomial model.

We get (using, for example, the Lagrangian bases)

$$y = -0.0138x^6 + 0.5084x^5 - 6.4279x^4 + 34.8575x^3 - 73.9916x^2 + 64.3128x - 18.0951.$$

Note that, the function changes from increasing to decreasing.

Therefore, this model function may not give good predictions.



Example 2

Consider the data set:

x_i	0.2	0.3	0.4	0.6	0.9
Case 1: y_i	2.7536	3.2411	3.8016	5.1536	7.8671
Case 2: y_i	2.7536	3.2411	3.8916	5.1536	7.8671

We consider fitting the data by a 4-th order polynomial:

$$P_4(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4.$$

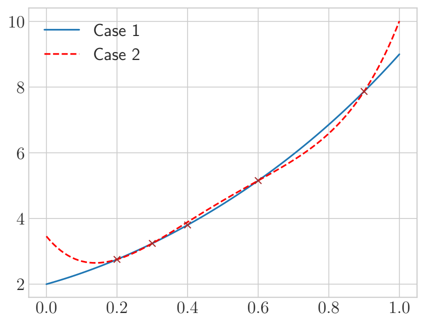
We assume that Case 1 gives the exact data.

In Case 2, we assume there is a measurement error at $x_i = 0.4$.

The results are shown in the following table.

	a_0	a_1	a_2	a_3	a_4
Case 1	2	3	4	-1	1
Case 2	3.4580	-13.2000	64.7500	-91.0000	46.0000

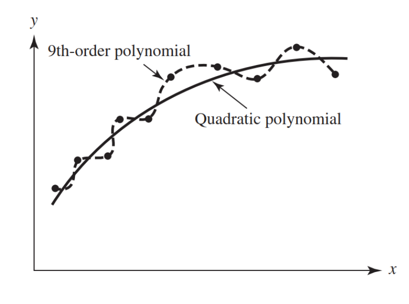
Thus, a small error in the data gives a completely different solution.



Smoothing

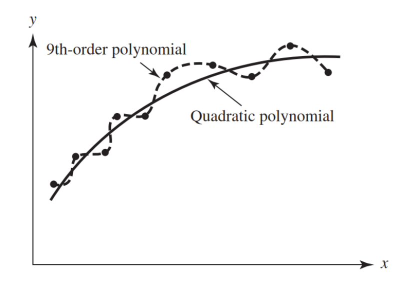
Recall that, high-order polynomials give too many oscillations and are sensitive to errors.

We introduce smoothing, which is a technique of using lower-order polynomials to capture the trend in the data.



Note:

- Using a 9-th order polynomial (10 data points) gives an oscillatory model function.
- Using a lower-order polynomial (quadratic in this case) gives a smoother model function which can still capture the trend.
- The lower-order polynomial does not necessarily pass through all data points.



Two decisions of smoothing

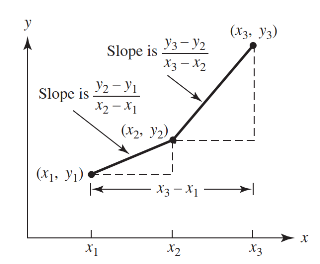
The process of smoothing requires two decisions:

1. the order of the interpolating polynomial must be selected—
    - we discuss this now,
    - the main tool is using divided differences;
2. the coefficients of the polynomial must be determined—
    - one uses the methods introduced in Chap. 3, since the type of the model function has been determined,
    - e.g. the least-squares criterion.

Divided differences

Consider the data points $(x_1, y_1), (x_2, y_2)$ and (x_3, y_3).

- $\frac{y_2 - y_1}{x_2 - x_1}$ can be regarded as an approximation to the first derivative over $[x_1, x_2]$,
- $\frac{y_3 - y_2}{x_3 - x_2}$ can be regarded as an approximation to the first derivative over $[x_2, x_3]$.



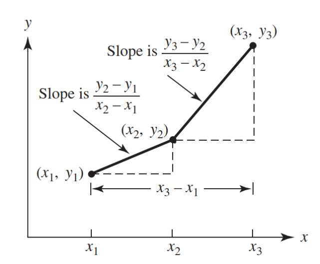
These are called first divided differences.

How about second derivatives (the derivative of the first derivative)?

One can use the number

$$\frac{\frac{y_3-y_2}{x_3-x_2}-\frac{y_2-y_1}{x_2-x_1}}{x_3-x_1}$$

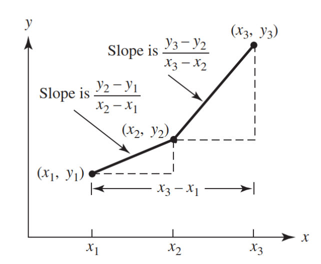
as an approximation to the second derivative over the interval $[x_1, x_3]$.



This is called a second divided difference.

We obtain the following table, called the divided difference table.

Data		First divided difference	Second divided difference
x_1	y_1		
		$\dfrac{y_2 - y_1}{x_2 - x_1}$	
			$\dfrac{\dfrac{y_3 - y_2}{x_3 - x_2} - \dfrac{y_2 - y_1}{x_2 - x_1}}{x_3 - x_1}$
x_2	y_2		
		$\dfrac{y_3 - y_2}{x_3 - x_2}$	
x_3	y_3		



General rule: Assume n-th divided differences are obtained. To get $(n+1)$-th divided differences, we take the difference between adjacent n-th divided differences and then divide it by the length of the interval over which the change has taken place.

An example

Consider the data set:

x_i	0	2	4	6	8
y_i	0	4	16	36	64

We obtain the following divided difference table:

	Data		Divided differences		
	x_i	y_i	Δ	Δ^2	Δ^3
	0	0			
			$4/2 = 2$		
	2	4		$4/4 = 1$	
			$12/2 = 6$		$0/6 = 0$
$\Delta x = 6$	4	16		$4/4 = 1$	
			$20/2 = 10$		$0/6 = 0$
	6	36		$4/4 = 1$	
			$28/2 = 14$		
	8	64			

Example: tape recorder (revisited)

Consider the data set

c_i	100	200	300	400	500	600	700	800
t_i (sec)	205	430	677	945	1233	1542	1872	2224

We have already constructed a 7th-order polynomial model.

We will now construct a lower-order polynomial model.

Two steps:

- determine the order of the polynomial;
- find the coefficients in the polynomial.

Step 1: We need divided differences. We obtain the following divided difference table:

Data		Divided differences			
x_i	y_i	$\underline{\Delta}$	$\underline{\Delta}^2$	$\underline{\Delta}^3$	$\underline{\Delta}^4$
100	205				
		2.2500			
200	430		0.0011		
		2.4700		0.0000	
300	677		0.0011		0.0000
		2.6800		0.0000	
400	945		0.0010		0.0000
		2.8800		0.0000	
500	1233		0.0011		0.0000
		3.0900		0.0000	
600	1542		0.0011		0.0000
		3.3000		0.0000	
700	1872		0.0011		
		3.5200			
800	2224				

From the table, we see the third divided differences are almost zero.

Hence, it is reasonable to assume that a quadratic polynomial will fit the data well.

Step 2: We will fit a quadratic polynomial $P(c) = a + bc + dc^2$.

We use the least-squares criterion:

$$S(a, b, d) = \sum_{i=1}^{m} |t_i - (a + bc_i + dc_i^2)|^2.$$

Taking partial derivatives,

$$\begin{aligned}
0 &= \frac{\partial S}{\partial a} = \sum_{i=1}^{m} (-2)(t_i - a - bc_i - dc_i^2), \\
0 &= \frac{\partial S}{\partial b} = \sum_{i=1}^{m} (-2c_i)(t_i - a - bc_i - dc_i^2), \\
0 &= \frac{\partial S}{\partial d} = \sum_{i=1}^{m} (-2c_i^2)(t_i - a - bc_i - dc_i^2).
\end{aligned}$$

Hence, we obtain the following system:

$$
\begin{aligned}
a(\sum_{i=1}^{m} 1) + b(\sum_{i=1}^{m} c_i) + d(\sum_{i=1}^{m} c_i^2) &= \sum_{i=1}^{m} t_i, \\
a(\sum_{i=1}^{m} c_i) + b(\sum_{i=1}^{m} c_i^2) + d(\sum_{i=1}^{m} c_i^3) &= \sum_{i=1}^{m} c_i t_i, \\
a(\sum_{i=1}^{m} c_i^2) + b(\sum_{i=1}^{m} c_i^3) + d(\sum_{i=1}^{m} c_i^4) &= \sum_{i=1}^{m} c_i^2 t_i,
\end{aligned}
$$

where c_i and t_i are obtained from the table:

c_i	100	200	300	400	500	600	700	800
t_i (sec)	205	430	677	945	1233	1542	1872	2224

Using the data, we have

$$
\begin{aligned}
8a + 36b + 204d &= 9128, \\
36a + 204b + 1296d &= 53{,}189, \\
204a + 1296b + 8772d &= 343{,}539.
\end{aligned}
$$

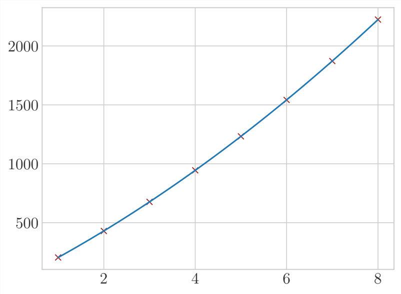
Solving it, we have

$$a = 0.142, \quad b = 194.226, \quad d = 10.464.$$

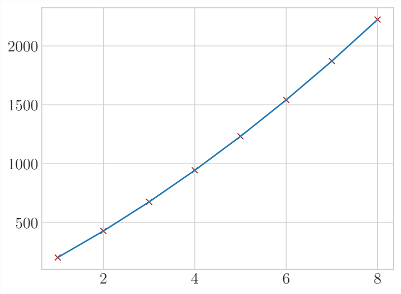
Thus, the model function is

$$P(c) = 0.142 + 194.226c + 10.464c^2.$$

We see that a lower-order polynomial can effectively capture the trend.



Lower-order model



High-order model

Example: stopping distance

Problem: Determine the stopping distance as a function of the speed of the car.

The following data set is obtained.

Speed v (mph)	20	25	30	35	40	45	50	55	60	65	70	75	80
Distance d (ft)	42	56	73.5	91.5	116	142.5	173	209.5	248	292.5	343	401	464

We will construct a model using a lower-order polynomial.

Step 1: Construct a divided difference table.

Data		Divided differences			
v_i	d_i	Δ	Δ^2	Δ^3	Δ^4
20	42				
		2.2800			
25	56		0.0700		
		3.5000		−0.0040	
30	73.5		0.0100		0.0006
		3.6000		0.0080	
35	91.5		0.1300		−0.0007
		4.9000		−0.0060	
40	116		0.0400		0.0004
		5.3000		0.0027	
45	142.5		0.0800		0.0000
		6.1000		0.0027	
50	173		0.1200		−0.0004
		7.3000		−0.0053	
55	209.5		0.0400		0.0005
		7.7000		0.0053	
60	248		0.1200		−0.0003
		8.9000		0.0000	
65	292.5		0.1200		0.0001
		10.1000		0.0020	
70	343		0.1500		−0.0003
		11.6000		−0.0033	
75	401		0.1000		
		12.6000			
80	464				

Note: 3-rd divided differences are small compared to first and second divided differences.

We will, again, find a quadratic model $P(v) = a + bv + cv^2$.

Step 2: Similar to the previous example, we obtain the following system:

$$
\begin{aligned}
a(\sum_{i=1}^{m} 1) + b(\sum_{i=1}^{m} v_i) + c(\sum_{i=1}^{m} v_i^2) &= \sum_{i=1}^{m} d_i, \\
a(\sum_{i=1}^{m} v_i) + b(\sum_{i=1}^{m} v_i^2) + c(\sum_{i=1}^{m} v_i^3) &= \sum_{i=1}^{m} v_i d_i, \\
a(\sum_{i=1}^{m} v_i^2) + b(\sum_{i=1}^{m} v_i^3) + c(\sum_{i=1}^{m} v_i^4) &= \sum_{i=1}^{m} v_i^2 d_i,
\end{aligned}
$$

where v_i and d_i are obtained from the data set:

Speed v (mph)	20	25	30	35	40	45	50	55	60	65	70	75	80
Distance d (ft)	42	56	73.5	91.5	116	142.5	173	209.5	248	292.5	343	401	464

Using the data, we have

$$\begin{aligned}
13\,a + 650\,b + 37050\,c &= 2652.5,\\
650\,a + 37050\,b + 2307500\,c &= 163970,\\
37050\,a + 2307500\,b + 152343750\,c &= 10804975.
\end{aligned}$$

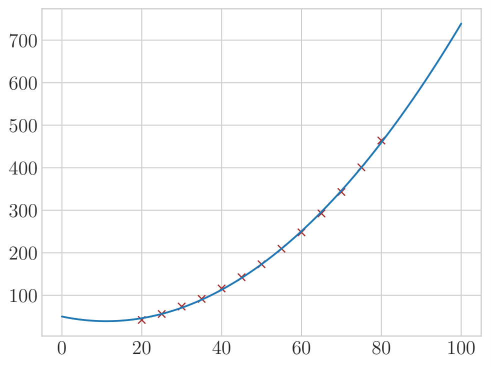
Solving it, we have

$$a = 50.0594, \quad b = -1.9701, \quad c = 0.0886.$$

Thus, the model function is

$$P(v) = 50.0594 - 1.9701v + 0.0886v^2.$$

We obtained a good model: $P(v) = 50.0594 - 1.9701v + 0.0886v^2$.



Cubic spline model

We discuss cubic spline models in this section.

Key idea:

- Focus locally first.
- Use local low-order polynomials.
- Connect the low-order polynomials to obtain the global fitted curve.

What is a cubic spline?

It is a cubic polynomial between successive data points.

Cubic spline: A function that is a cubic polynomial between successive data points.

Consider data points: $(x_1, y_1), (x_2, y_2)$ and (x_3, y_3).

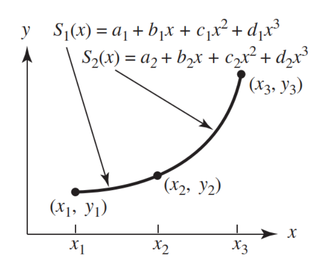
The cubic spline $S(x)$ is

- a cubic polynomial on $[x_1, x_2]$

$$S_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3,$$

- a cubic polynomial on $[x_2, x_3]$

$$S_2(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3.$$



Q: How do we find $S(x)$?

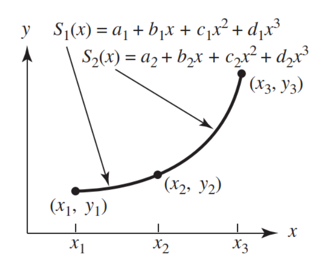
The following conditions are required for finding $S(x)$. Note that we need 8 conditions.

- $S(x)$ goes through data points.
  On the interval $[x_1, x_2]$:

$$y_1 = S_1(x_1) = a_1 + b_1 x_1 + c_1 x_1^2 + d_1 x_1^3,$$
$$y_2 = S_1(x_2) = a_1 + b_1 x_2 + c_1 x_2^2 + d_1 x_2^3.$$

  On the interval $[x_2, x_3]$:

$$y_2 = S_2(x_2) = a_2 + b_2 x_2 + c_2 x_2^2 + d_2 x_2^3,$$
$$y_3 = S_2(x_3) = a_2 + b_2 x_3 + c_2 x_3^2 + d_2 x_3^3.$$



Remark

There are 4 conditions.

- $S'(x)$ is continuous at interior data points

$$S_1'(x) = b_1 + 2c_1x + 3d_1x^2,$$
$$S_2'(x) = b_2 + 2c_2x + 3d_2x^2.$$

Continuity at x_2:

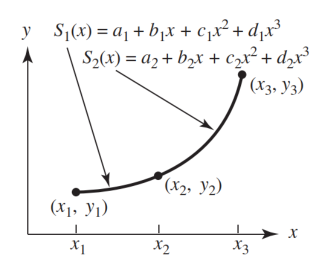
$$b_1 + 2c_1x_2 + 3d_1x_2^2 = b_2 + 2c_2x_2 + 3d_2x_2^2.$$

- $S''(x)$ is continuous at interior data points

$$S_1''(x) = 2c_1 + 6d_1x,$$
$$S_2''(x) = 2c_2 + 6d_2x.$$

Continuity at x_2:

$$2c_1 + 6d_1x_2 = 2c_2 + 6d_2x_2.$$



Remark

We have 2 more conditions.

Finally, we need 2 extra conditions.

The following choice gives the natural cubic spline.

- $S''(x) = 0$ at the two end-points

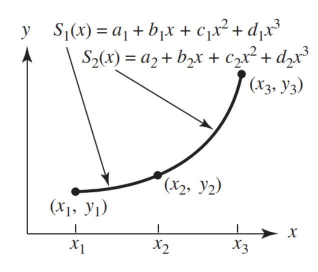
$$S_1''(x) = 2c_1 + 6d_1 x,$$
$$S_2''(x) = 2c_2 + 6d_2 x.$$

At x_1:

$$2c_1 + 6d_1 x_1 = 0.$$

At x_3:

$$2c_2 + 6d_2 x_3 = 0.$$



Remark

The last 2 conditions.

An example

Consider the data set:

x	1	2	3
y	5	8	25

We first write down the equations.

- $S(x)$ goes through data points:
  On the interval $[1, 2]$:

$$5 = S_1(1) = a_1 + b_1(1) + c_1(1)^2 + d_1(1)^3,$$
$$8 = S_1(2) = a_1 + b_1(2) + c_1(2)^2 + d_1(2)^3.$$

  On the interval $[2, 3]$:

$$8 = S_2(2) = a_2 + b_2(2) + c_2(2)^2 + d_2(2)^3,$$
$$25 = S_2(3) = a_2 + b_2(3) + c_2(3)^2 + d_2(3)^3.$$

x	1	2	3
y	5	8	25

- $S'(x)$ is continuous at interior data points:

$$b_1 + 2c_1(2) + 3d_1(2)^2 = b_2 + 2c_2(2) + 3d_2(2)^2.$$

- $S''(x)$ is continuous at interior data points:

$$2c_1 + 6d_1(2) = 2c_2 + 6d_2(2).$$

- $S''(x) = 0$ at the two end-points
  At x_1:

$$2c_1 + 6d_1(1) = 0,$$

  At x_3:

$$2c_2 + 6d_2(3) = 0.$$

The idea is to first solve c_1, d_1, c_2, d_2 in terms of b_1, b_2.

From the last four equations, we have

$$c_1 = \frac{b_2 - b_1}{8}, \qquad d_1 = \frac{b_1 - b_2}{24},$$
$$c_2 = \frac{3(b_1 - b_2)}{8}, \qquad d_2 = \frac{b_2 - b_1}{24}.$$

Using these in the first 4 equations,

$$5 = a_1 + b_1 + \frac{b_2 - b_1}{8} + \frac{b_1 - b_2}{24},$$
$$8 = a_1 + 2b_1 + \frac{b_2 - b_1}{2} + \frac{b_1 - b_2}{3},$$
$$8 = a_2 + 2b_2 + \frac{3(b_1 - b_2)}{2} + \frac{b_2 - b_1}{3},$$
$$25 = a_2 + 3b_2 + \frac{27(b_1 - b_2)}{8} + \frac{9(b_2 - b_1)}{8}.$$

Eliminating a_1 and a_2, we get

$$3 = \frac{11b_1 + b_2}{12}, \qquad 17 = \frac{13b_1 - b_2}{12}.$$

Solving, we get

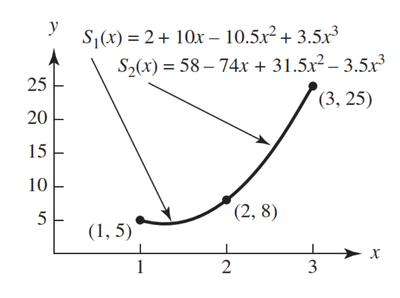
$$b_1 = 10, \qquad b_2 = -74.$$

The other six unknowns can be solved easily

$$a_1 = 2, a_2 = 58, \quad c_1 = -10.5, c_2 = 31.5, \quad d_1 = 3.5, d_2 = -3.5.$$

Hence the cubic spline $S(x)$ is

$$S_1(x) = 2 + 10x - 10.5x^2 + 3.5x^3, \quad x \in [1, 2],$$
$$S_2(x) = 58 - 74x + 31.5x^2 - 3.5x^3, \quad x \in [2, 3].$$



$$S_1(x) = 2 + 10x - 10.5x^2 + 3.5x^3, \quad x \in [1, 2],$$
$$S_2(x) = 58 - 74x + 31.5x^2 - 3.5x^3, \quad x \in [2, 3].$$

For example, if we need to predict the value at $x = 1.67$, we can evaluate $S(1.67)$.

Since $1.67 \in [1, 2]$, we have $S(1.67) = S_1(1.67) = 5.72$.

Generalization

The construction of cubic spline can be generalized.

Let (x_i, y_i), $i = 1, 2, \ldots, m+1$ be a set of data points.

The cubic spline $S(x)$ is a cubic polynomial on each $[x_i, x_{i+1}]$,

$$S(x) = \begin{cases} S_1(x) & = a_1 + b_1 x + c_1 x^2 + d_1 x^3, \quad x \in [x_1, x_2], \\ S_2(x) & = a_2 + b_2 x + c_2 x^2 + d_2 x^3, \quad x \in [x_2, x_3], \\ & \vdots \\ S_m(x) & = a_m + b_m x + c_m x^2 + d_m x^3, \quad x \in [x_m, x_{m+1}]. \end{cases}$$

We need $4m$ equations.

First, $S(x)$ goes through all data points.

On $[x_1, x_2]$,

$$y_1 = S_1(x_1) = a_1 + b_1x_1 + c_1x_1^2 + d_1x_1^3,$$
$$y_2 = S_1(x_2) = a_1 + b_1x_2 + c_1x_2^2 + d_1x_2^3.$$

On $[x_2, x_3]$,

$$y_2 = S_2(x_2) = a_2 + b_2x_2 + c_2x_2^2 + d_2x_2^3,$$
$$y_3 = S_2(x_3) = a_2 + b_2x_3 + c_2x_3^2 + d_2x_3^3.$$

On $[x_m, x_{m+1}]$,

$$y_m = S_m(x_m) = a_m + b_mx_m + c_mx_m^2 + d_mx_m^3,$$
$$y_{m+1} = S_m(x_{m+1}) = a_m + b_mx_{m+1} + c_mx_{m+1}^2 + d_mx_{m+1}^3.$$

There are $2m$ equations.

Second, $S'(x)$ is continuous at interior points.

At x_2, we need $S_1'(x_2) = S_2'(x_2)$:

$$b_1 + 2c_1x_2 + 3d_1x_2^2 = b_2 + 2c_2x_2 + 3d_2x_2^2.$$

At x_3, we need $S_2'(x_3) = S_3'(x_3)$:

$$b_2 + 2c_2x_3 + 3d_2x_3^2 = b_3 + 2c_3x_3 + 3d_3x_3^2.$$

At x_m, we need $S_{m-1}'(x_m) = S_m'(x_m)$:

$$b_{m-1} + 2c_{m-1}x_m + 3d_{m-1}x_m^2 = b_m + 2c_mx_m + 3d_mx_m^2.$$

There are $m-1$ equations.

Third, $S''(x)$ is continuous at interior points.

At x_2, we need $S_1''(x_2) = S_2''(x_2)$:

$$2c_1 + 6d_1x_2 = 2c_2 + 6d_2x_2.$$

At x_3, we need $S_2''(x_3) = S_3''(x_3)$:

$$2c_2 + 6d_2x_3 = 2c_3 + 6d_3x_3.$$

At x_m, we need $S_{m-1}''(x_m) = S_m''(x_m)$:

$$2c_{m-1} + 6d_{m-1}x_m = 2c_m + 6d_mx_m.$$

There are $m-1$ equations.

Finally, we add 2 more conditions at end-points,

$$S_1''(x_1) = 0, \qquad S_m''(x_{m+1}) = 0.$$

That is,

$$2c_1 + 6d_1x_1 = 0, \qquad 2c_m + 6d_mx_{m+1} = 0.$$

There are totally $4m$ equations.

We can determine all coefficients in $S(x)$.

One needs to write a computer code to solve this. For example, there is a built-in class `CubicSpline` in `scipy`—a famous python package—to do this, and you generally need to a few lines of codes.

A remark

The choice

$$S_1''(x_1) = 0, \qquad S_m''(x_{m+1}) = 0$$

gives smallest curvature. Note for a curve $(x, f(x))$, the mathematical definition of the curvature at x is $f''/(1+f'^2)^{3/2}$.

Let G be the cubic spline with other choices of $G''(x_1)$ and $G''(x_{m+1})$, then we have

$$\int_{x_1}^{x_{m+1}} \left(S''\right)^2 \mathrm{d}x \leq \int_{x_1}^{x_{m+1}} \left(G''\right)^2 \mathrm{d}x.$$

To show this

$$\int_{x_1}^{x_{m+1}} \left(G''\right)^2 \mathrm{d}x = \int_{x_1}^{x_{m+1}} \left(G'' - S'' + S''\right)^2 \mathrm{d}x$$

$$= \int_{x_1}^{x_{m+1}} \left(G'' - S''\right)^2 \mathrm{d}x + 2\int_{x_1}^{x_{m+1}} \left(G'' - S''\right)S'' \,\mathrm{d}x + \int_{x_1}^{x_{m+1}} \left(S''\right)^2 \mathrm{d}x.$$

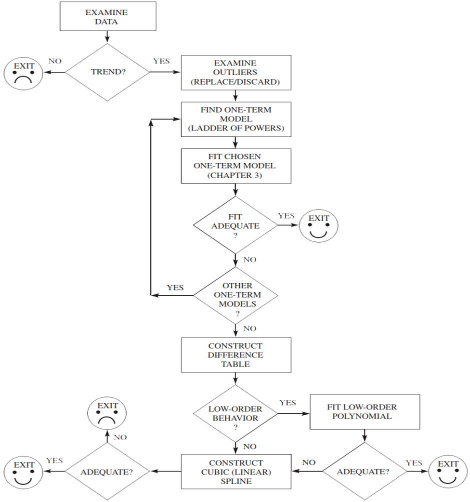
We will show

$$\int_{x_1}^{x_{m+1}} (G'' - S'')S'' \, \mathrm{d}x = 0.$$

Indeed,

$$\begin{aligned}
\int_{x_1}^{x_{m+1}} (G'' - S'')S'' \, \mathrm{d}x &= \sum_{i=1}^{m} \int_{x_i}^{x_{i+1}} (G'' - S'')S'' \, \mathrm{d}x \\
&= \sum_{i=1}^{m} \left\{ - \int_{x_i}^{x_{i+1}} (G' - S')S''' \, \mathrm{d}x + (G' - S')S'' \Big|_{x_i}^{x_{i+1}} \right\} \\
&= \sum_{i=1}^{m} \left\{ - \int_{x_i}^{x_{i+1}} (G' - S')S''' \, \mathrm{d}x \right\} \\
&= \sum_{i=1}^{m} \left\{ - S_i''' \Big((G - S)(x_{i+1}) - (G - S)(x_i) \Big) \right\} = 0.
\end{aligned}$$

Summary


EXAMINE DATA
TREND?
NO
EXIT
YES
EXAMINE OUTLIERS (REPLACE/DISCARD)
FIND ONE-TERM MODEL (LADDER OF POWERS)
FIT CHOSEN ONE-TERM MODEL (CHAPTER 3)
FIT ADEQUATE ?
YES
EXIT
NO
OTHER ONE-TERM MODELS ?
YES
NO
CONSTRUCT DIFFERENCE TABLE
LOW-ORDER BEHAVIOR ?
YES
FIT LOW-ORDER POLYNOMIAL
NO
ADEQUATE?
YES
EXIT
NO
CONSTRUCT CUBIC (LINEAR) SPLINE
ADEQUATE?
YES
EXIT
NO
EXIT

Disclaimer

All figures, tables, and data appearing in the slides are only used for teaching under guidelines of Fair Use.