Lecture 5 Empirical model building - Interpolation

Mathematical Modeling

Prof. Dr. Jingzhi Li

Department of Mathematics,
Southern University of Science and Technology

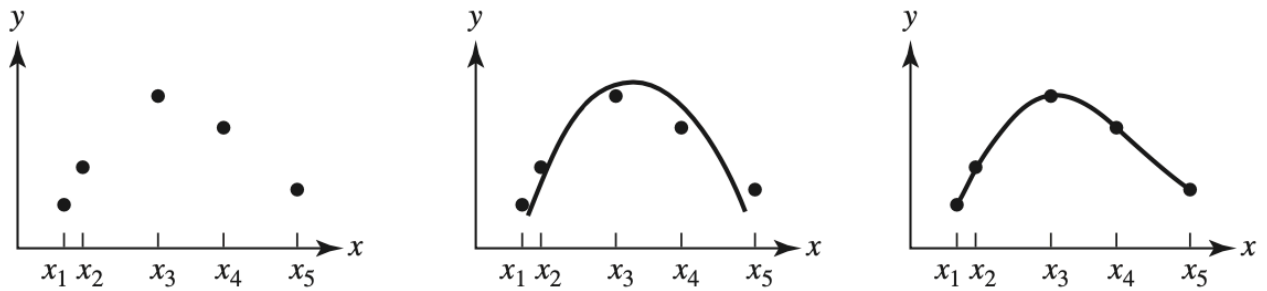
2025 Spring

Data and modeling: three situations

- **Given a data set we may take 3 approaches:**
  1. **Fit an already selected model type to the data**
     - The model type is already fixed
     - For example: a linear, or a quadratic model; mass-action model
     - $y = kx + b$
  2. **Choose the most appropriate model from several alternative models that have been fitted to the data**
     - Decide whether the best-fitting exponential model is better than the best-fitting polynomial model
     - $y = kx + b$, $y = ae^x$, $y = c \ln x$
  3. **Make predictions based solely on the data**
     - No hypothesis regarding the type of model
     - Predict intermediate and/or future behavior based just on the data set

Empirical model building

- Scenario for this lecture: the modeler has no clue regarding the mathematical form of the model (s)he is looking for
  - does not know what kind of curve describes the behavior
  - can only be guided by the data
- Aim: build an empirical model based on the data set
  - seek a curve that captures the trend of the data
  - aim to predict the behavior in-between the data points
- Example
  - given a data set
  - look for a quadratic curve?
  - look for a perfect fit?

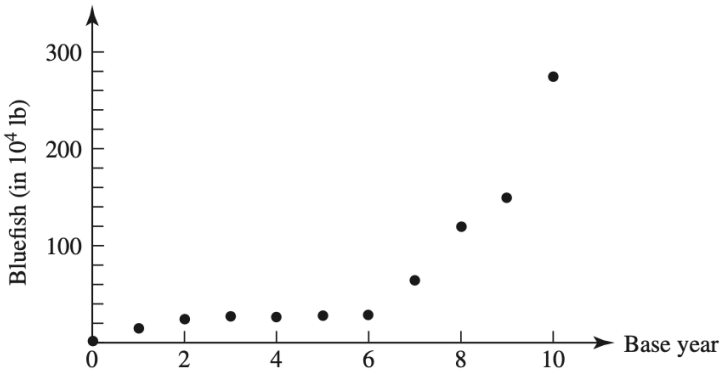


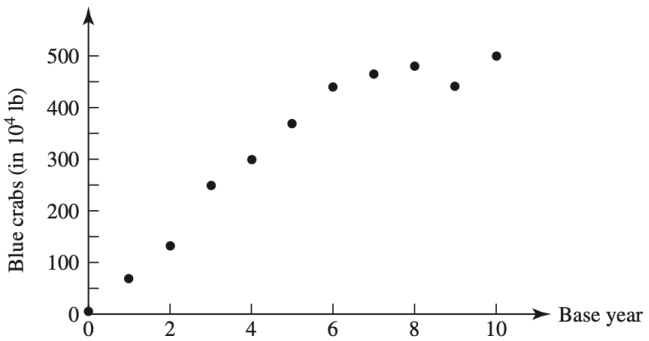
Example

- Given a data set on the harvesting of bluefish and blue crabs during 1940–1990 in Chesapeake Bay
- Aim: build a model accounting for the data
  - tendency to harvest more bluefish
  - suggests availability of bluefish

Table 4.1 Harvesting the bay, 1940–1990

Year	Bluefish (lb)	Blue crabs (lb)
1940	15,000	100,000
1945	150,000	850,000
1950	250,000	1,330,000
1955	275,000	2,500,000
1960	270,000	3,000,000
1965	280,000	3,700,000
1970	290,000	4,400,000
1975	650,000	4,660,000
1980	1,200,000	4,800,000
1985	1,500,000	4,420,000
1990	2,750,000	5,000,000





Example (continued)

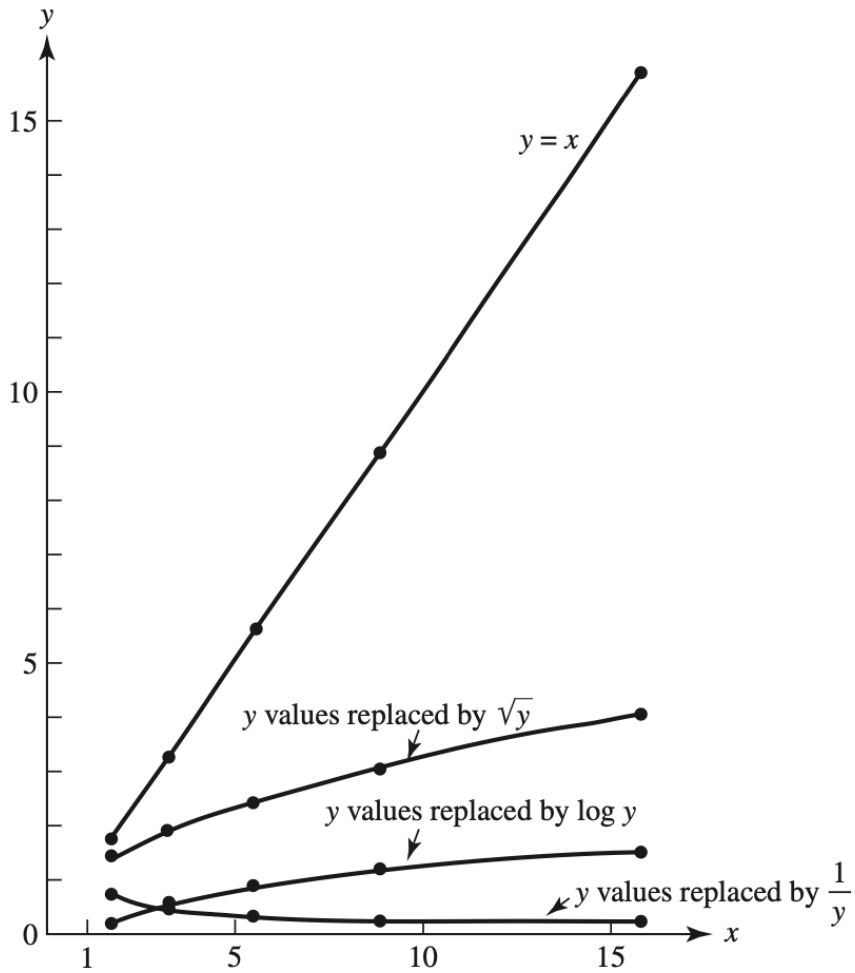
- Strategy: transform the data so that it visually resembles a linear model
  - which data transformation to choose?

- The ladder of powers:

Ladder of powers
$\vdots$
z^2
z
$\sqrt{z}$
$\log z$
$\frac{1}{\sqrt{z}}$
$\frac{1}{z}$
$\frac{1}{z^2}$
$\vdots$

Example (continued)

- Relative effects of three transformations

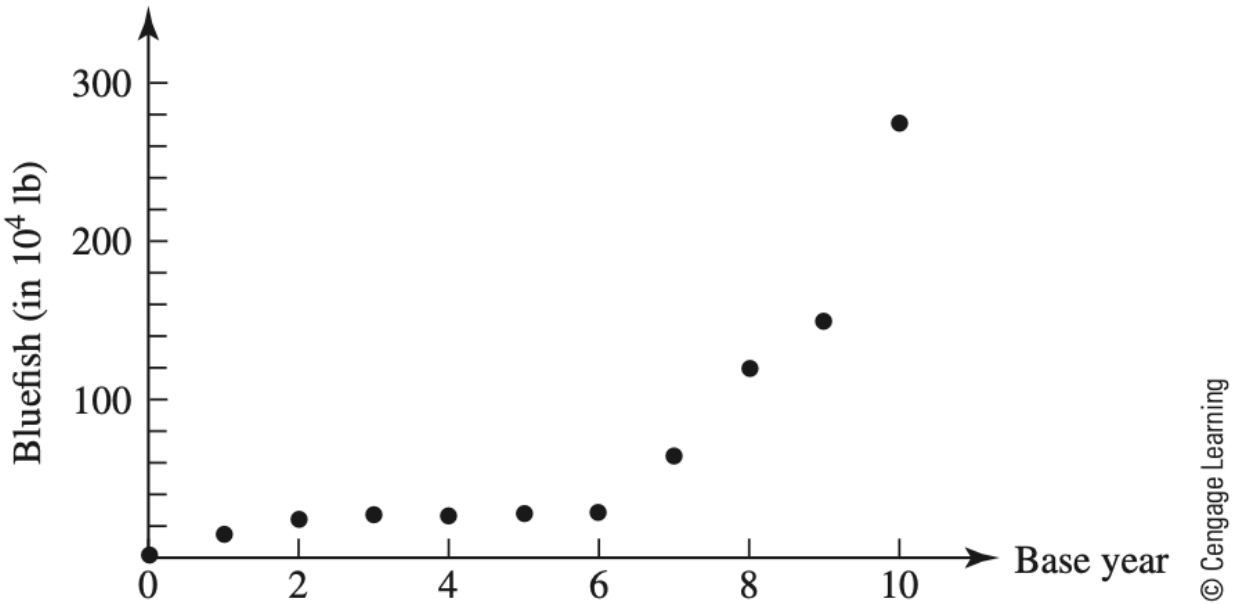


Example (continued)

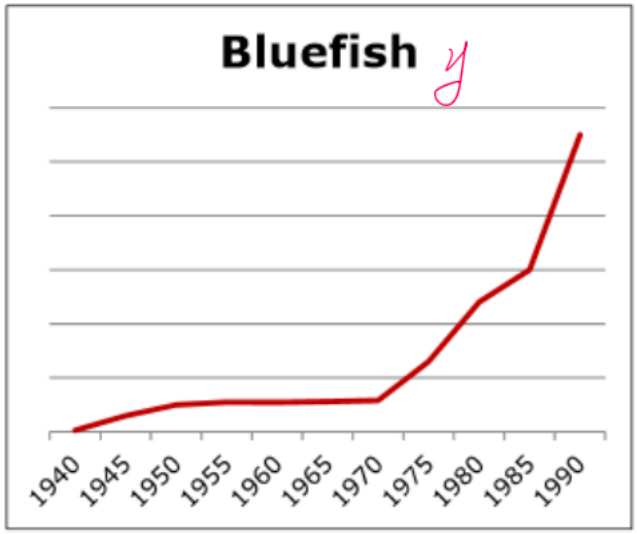
- When starting with a **convex** plot. To make it "more linear" :
  - squeeze the right-hand tail downward by changing the y values to $\sqrt{y}$, $\ln(y)$, or other choice from the ladder of powers
  - stretch the right-hand tail by replacing the x values with x^2, x^3, etc
- When starting with a **concave** plot. To make it "more linear" :
  - stretch the right-hand tail upward by changing the y values to y^2, y^3, etc
  - stretch the right-hand tail by replacing the x values with $\sqrt{x}$, $\ln(x)$, or a more drastic choice from the ladder of powers
- Note: a transformation of z into $1/z$, $1/z^2$, etc changes an increasing function into a decreasing one
  - often one adds a negative sign in front to maintain the monotonicity

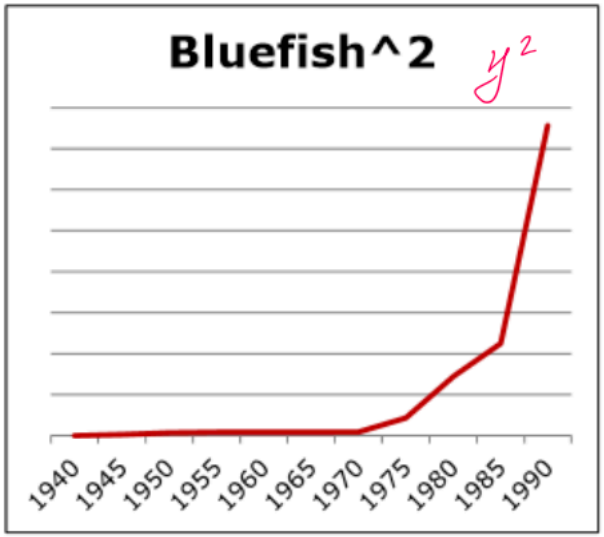
Example (continued)

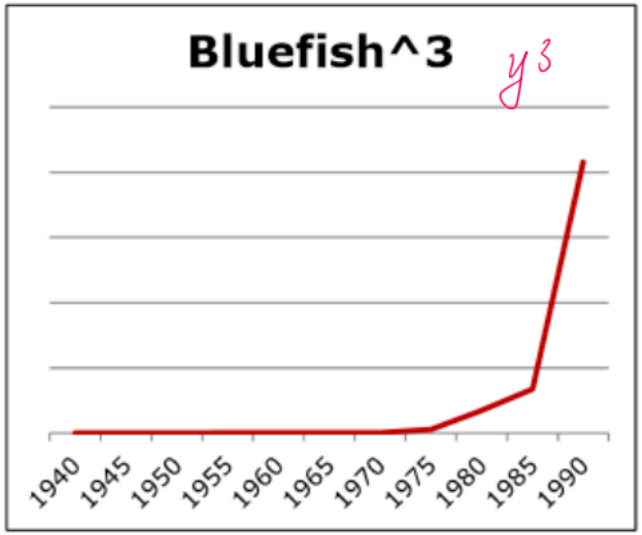
- Harvest of bluefish
  - linearize the data
  - changing x to other values upper in the ladder of powers: x^2, x^3, etc does not give a plot that "looks linear"
  - change y to lower values in the ladder: $\ln(y)$ works alright
  - See next slide

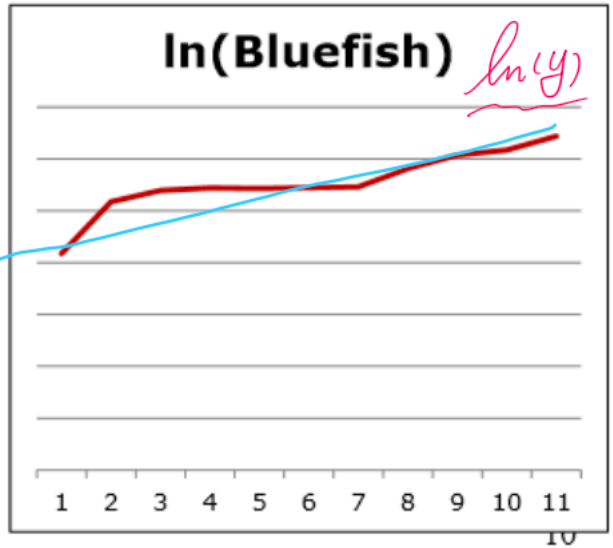


Example (continued



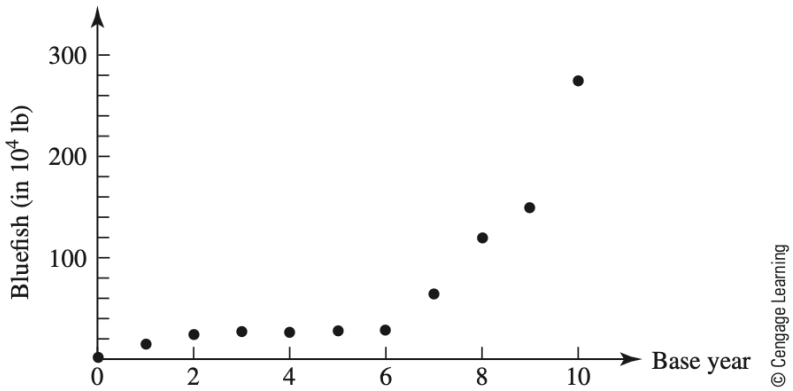


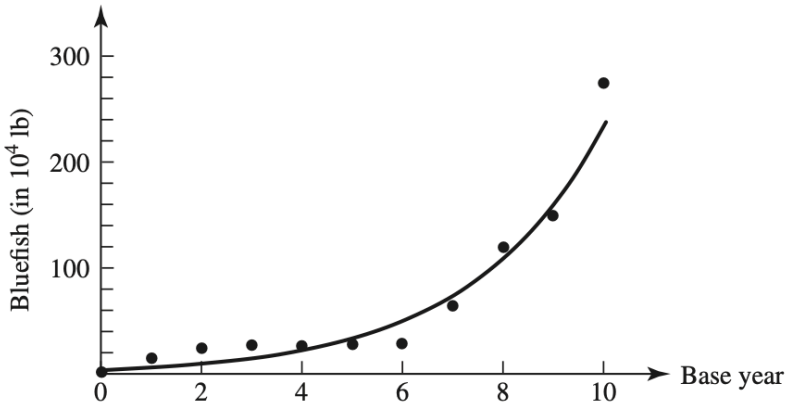




Example (continued)

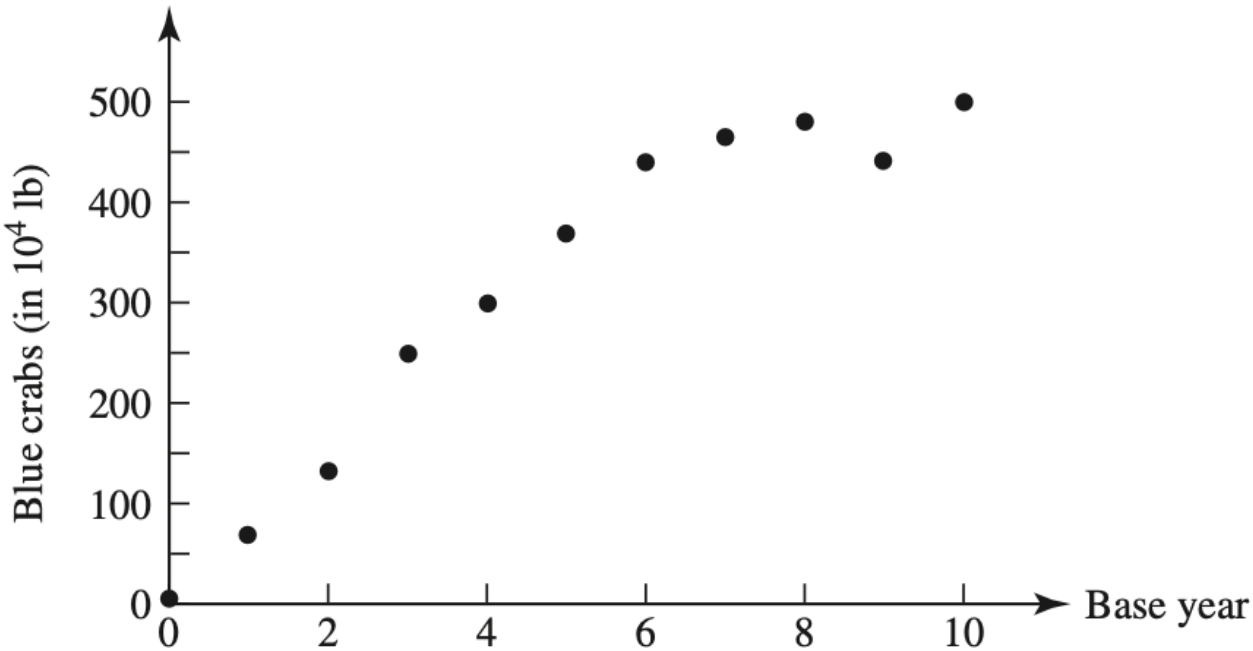
- Harvest of bluefish
  - linearize the data
  - choose the model $\ln(y) = mx + b$
  - fit the model (least-squares): $\ln(y) = 0.7231 + 0.1654x$
  - Final model: $y = 5.2857(1.4635)^x$
    - y measured in 10^4 pounds
    - x measured in offset wrt 1940



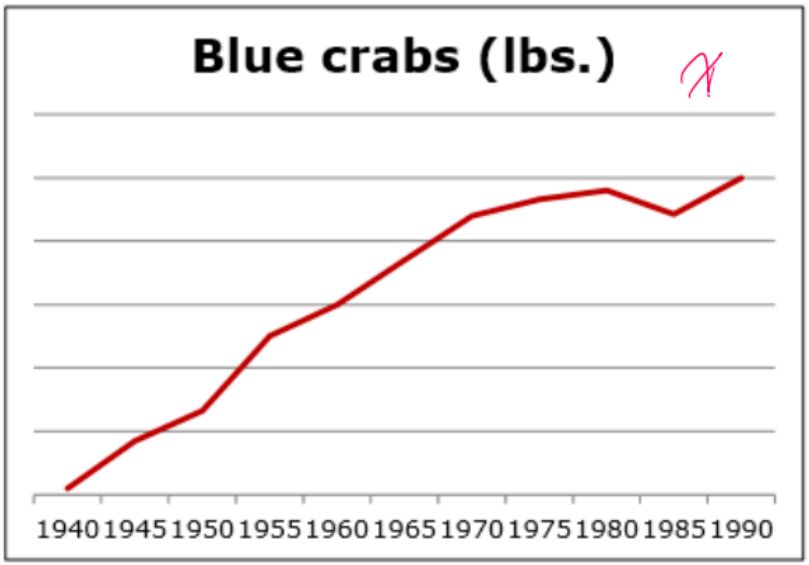


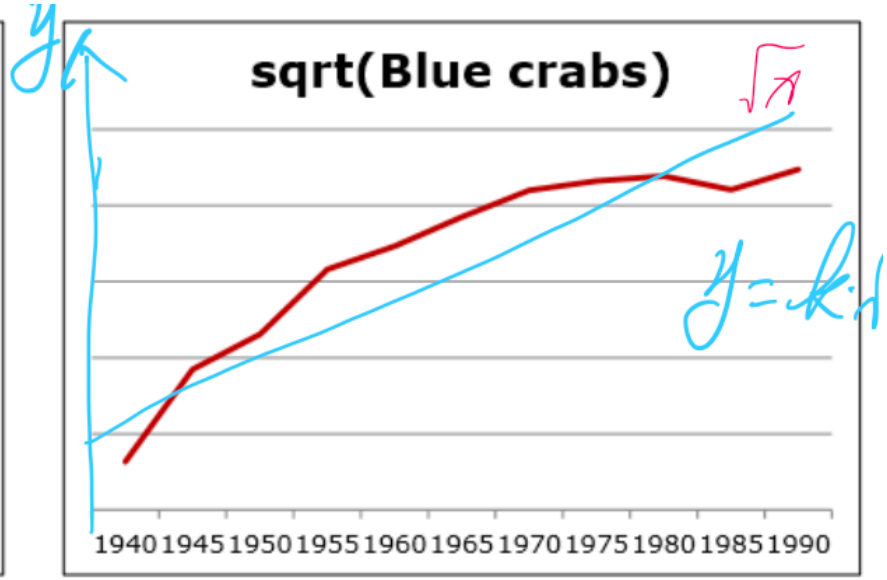
Example (continued)

- Harvest of blue crabs
  - linearize the data
  - choose to replace x with $\sqrt{x}$
    - squeezes the right-hand tail to the left



Example (continued

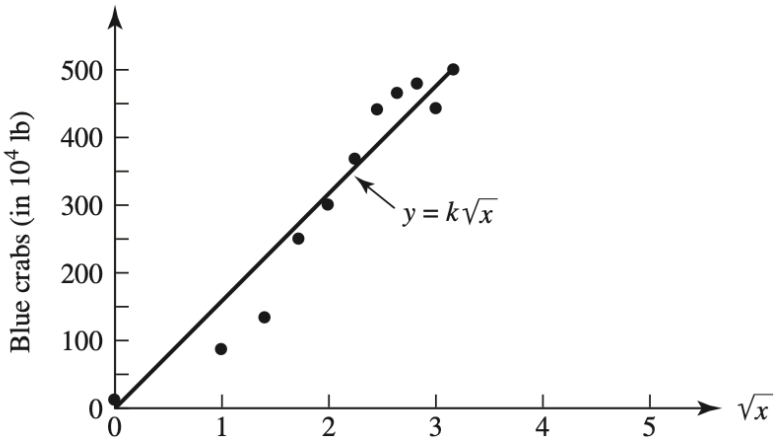




Example (continued)

- Harvest of blue crabs
  - linearize the data
  - choose to replace x with $\sqrt{x}$
  - choose the model $y = k\sqrt{x}$
  - fit the model (least-squares): $y = 158.344\sqrt{x}$
    - y measured in 10^4 pounds
    - x measured in offset wrt 1940

Figure 4.7
The line $y = 158.344\sqrt{x}$



Example (continued)

- Final models: how good are they?
  - pretty good with respect to the relative errors
  - not too good with respect to predictions for year 2010:
    - bluefish: 10.9 million pounds
    - blue crabs: 5.92 million pounds
- Note: better for predicting behavior in-between the data points than for extrapolating

Figure 4.5
Superimposed data and model $y = 5.2857(1.4635)^x$

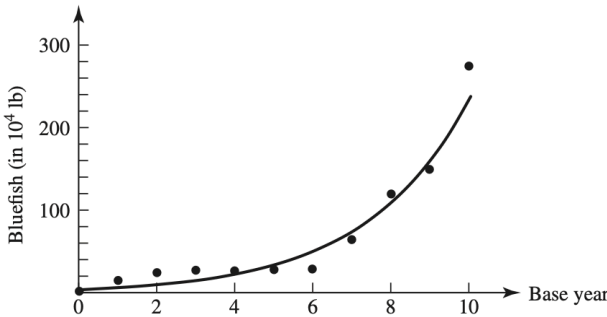
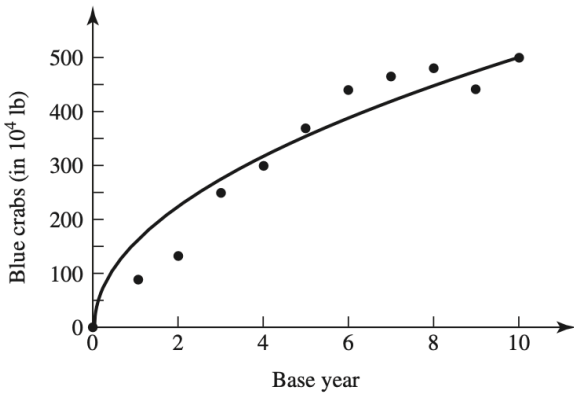


Figure 4.8
Superimposed data and model $y = 158.344\sqrt{x}$



Summary

- Analyze data to observe the trend
  - if obvious outliers are in the data, then double-check the point, or even eliminate the outlier
- Transform the data into an approximately linear plot
  - strictly qualitative judgment!
  - careful about being deceived by squeezing the data together
- Formulate the model based on the chosen data transformation
- Estimate the parameters
- Analyze the goodness of the fit
- This yields simple, one-term models

High-order polynomials

- When the simple, one-term models on the previous slides yield poor result: consider models with more than one term
  - in this lecture: consider polynomials as a model
- A main question in the remaining of this lecture: how to fit a polynomial to pass through a number of given points (the data set)

Example

- Given 3 data points, a unique polynomial of degree at most 2 passes through all 3 data points
  - an infinity of polynomials of higher degrees

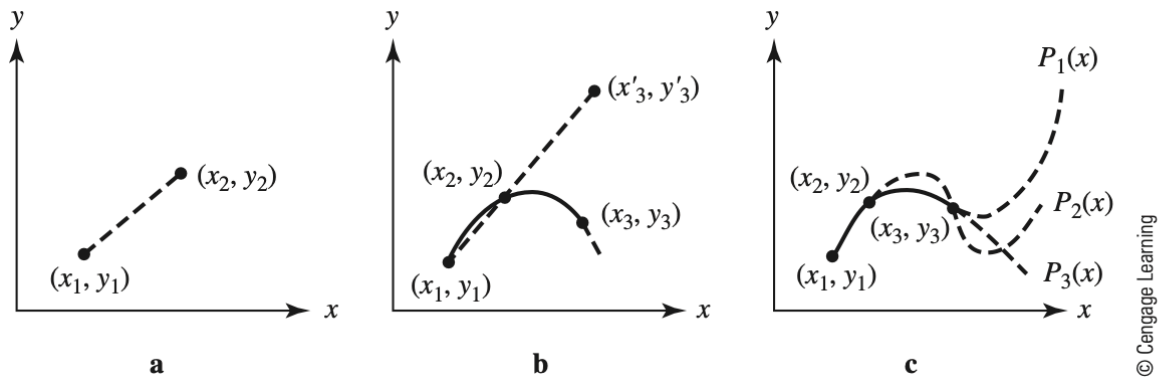


Figure 4.10

A unique polynomial of at most degree 2 can be passed through three data points (**a** and **b**), but an infinite number of polynomials of degree greater than 2 can be passed through three data points (**c**)

The Lagrangian form of a polynomial

- **Theorem (1779-Waring, 1783-Euler):**
  For a given sequence of pairs of real numbers $(x_0, y_0), (x_1, y_1), ..., (x_n, y_n)$, with $x_0, x_1, ..., x_n$ pair-wise distinct, there exists a unique polynomial $P(x)$ of degree at most n such that

$$y_k = P(x_k) \quad \text{for all } 0 \leq k \leq n$$

  Moreover, this polynomial can be written (in its Lagrangian form) as follows:

$$P(x) = y_0 L_0(x) + \ldots + y_n L_n(x)$$

  where

$$L_k(x) = \frac{(x - x_0)(x - x_1)\ldots(x - x_{k-1})(x - x_{k+1})\ldots(x - x_n)}{(x_k - x_0)(x_k - x_1)\ldots(x_k - x_{k-1})(x_k - x_{k+1})\ldots(x_k - x_n)}$$

Proof

- The polynomial above is a solution to the problem:

$$P(x_k) = y_k L_k(x_k) = y_k.$$

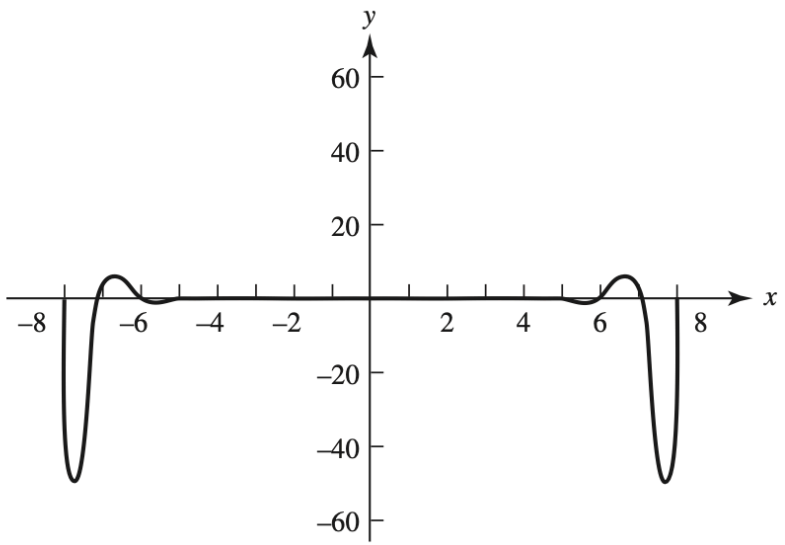
  - Indeed, because $L_i(x_j) = 0$ for all $i \neq j$.
- Prove the uniqueness:
  - Assume another such polynomial $Q(x)$ exists. Take $R(x) = P(x) - Q(x)$.
  - R has degree at most n (because both P and Q are such).
  - R has $n + 1$ distinct roots: $x_0, x_1, \ldots, x_n$.
  - By the Factor Theorem, $(x - x_0)(x - x_1) \cdots (x - x_n)$ is a factor of $R(x)$.
  - Based on the argument of the degree of R we get that $R(x) = 0$.

Observations

- A Lagrangian-form polynomial will always give perfect fit for any experimental data
  - so what if it has high-degree?
  - degree at most n for $n+1$ data points

- Problems with high-order polynomials
  - severe oscillations
    - see example on the right, where the data is (-8,0), (-6,0), ..., (8,0)
    - polynomial of degree at most 16 is unique: 0
    - a polynomial of degree higher than 16 showed in the figure
  - very sensitive to the experimental data
    - recalculate when data changes

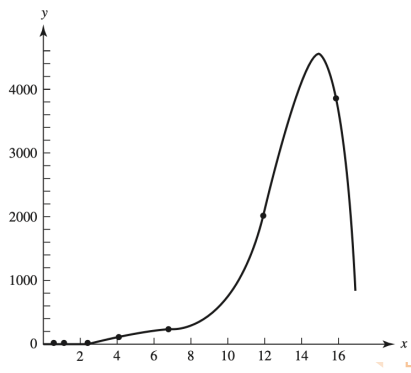
Figure 4.12
Fitting a higher-order polynomial through the data points in Table 4.10



Smoothing: low-order polynomials

- Aim: find a method that retains the advantages of working with polynomials but eliminates the disadvantages of high-order polynomials
- Popular technique: smoothing
  - choose a low-order polynomial regardless of the number of data points
  - In general we will have more data points than parameters
  - The low-order polynomial will not pass through all the points
  - Decide which low-order polynomial best fits the data (according to criteria discussed in the first half of this lecture)
  - Achieve less oscillations and less sensitivity to the data

Figure 4.14
The plot of 6th-order polynomial fit superimposed on the scatterplot



Smoothing: low-order polynomials

- Two decision to make
  - The degree of the interpolating polynomial
  - The coefficients of the polynomial
- This part is similar to the issues discussed in the first part of this lecture

Divided differences

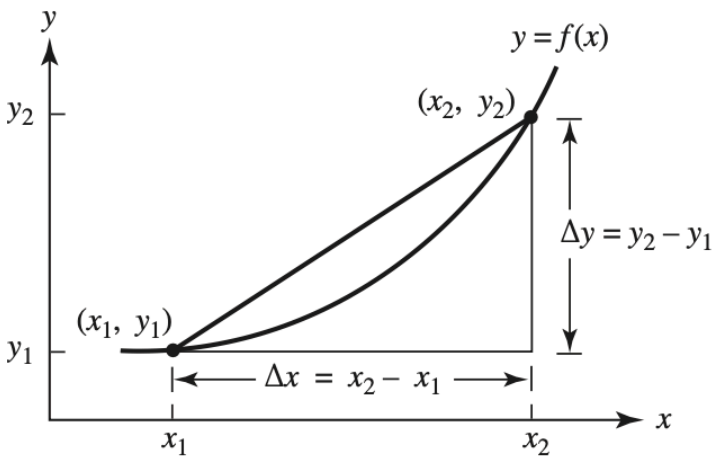
- Question: how do we choose the degree of the interpolating polynomial?
- Note:
  - A quadratic function $P(x) = a + bx + cx^2$ is characterized by having a constant second derivative and a zero third derivative
  - $P'(x) = b + 2cx$
  - $P''(x) = 2c$
  - $P'''(x) = 0$

- We only have info on the set of discrete data points
- Recall that:

$$\frac{dy}{dx}(x_1) = \lim_{x_2 \to x_1} \frac{y(x_2) - y(x_1)}{x_2 - x_1} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$$

Figure 4.17

The derivative of $y = f(x)$ at $x = x_1$ is the limit of the slope of the secant line.



Divided differences

- Recall that

$$\frac{dy}{dx}(x_1) = \lim_{x_2 \to x_1} \frac{y(x_2) - y(x_1)}{x_2 - x_1} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$$

- Given (x_i, y_i) we can calculate $(y_{i+1} - y_i, x_{i+1} - x_i)$
  - Unless the differences in the x-points are really small, not good estimates of the first derivative
  - However, if the derivative is to be zero (or constant), then $y_{i+1} - y_i$ should also be 0
- Repeat the procedure of calculating the differences using the new set of data to estimate the second derivative

Example: quadratic data

Table 4.14 A hypothetical set of collected data

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

Table 4.15 A difference table for the data of Table 4.14

Data		Differences			
x_i	y_i	Δ	Δ^2	Δ^3	Δ^4
0	0				
		4			
2	4		8		
		12		0	
4	16		8		0
		20		0	
6	36		8		
		28			
8	64				

Divided differences

- Note: because of the various errors or that data is not exactly quadratic, we may expect the differences not to be exactly zero, but just small
  - How small is small enough?
  - Calculate $(y_{i+1} - y_i)/(x_{i+1} - x_i)$

Table 4.16 The first and second divided differences estimate the first and second derivatives, respectively

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

Divided differences

- To see which x-points to consider for higher-order differences, draw diagonals in the table as shown in the example on the right

Table 4.17 A divided difference table for the data of Table 4.14

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

Example

- The calculations on the right show the data to be quadratic

Table 4.18 A divided difference table for the tape recorder data

Data		Divided differences			
x_i	y_i	Δ	Δ^2	Δ^3	Δ^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				

Observations on the difference tables

- Careful in the judgment of what is small and what is not
  - The measuring units!
  - Relative, qualitative judgment
- Careful about measurement errors
  - Measurement errors may yield the values in a column to be very close but not exactly equal
  - See example on the right, where the n-1 column should have been constant if not for the errors
  - start having a mix of positive and negative terms in a column

Δ^{n-1}	Δ^n	Δ^{n+1}
6.01		
	−0.01	
6.00		0.02
	0.01	
6.01		

Example

- The third divided differences are very small in magnitude when compared to the original data
- Negative signs have started to appear
- Decide that the data is quadratic

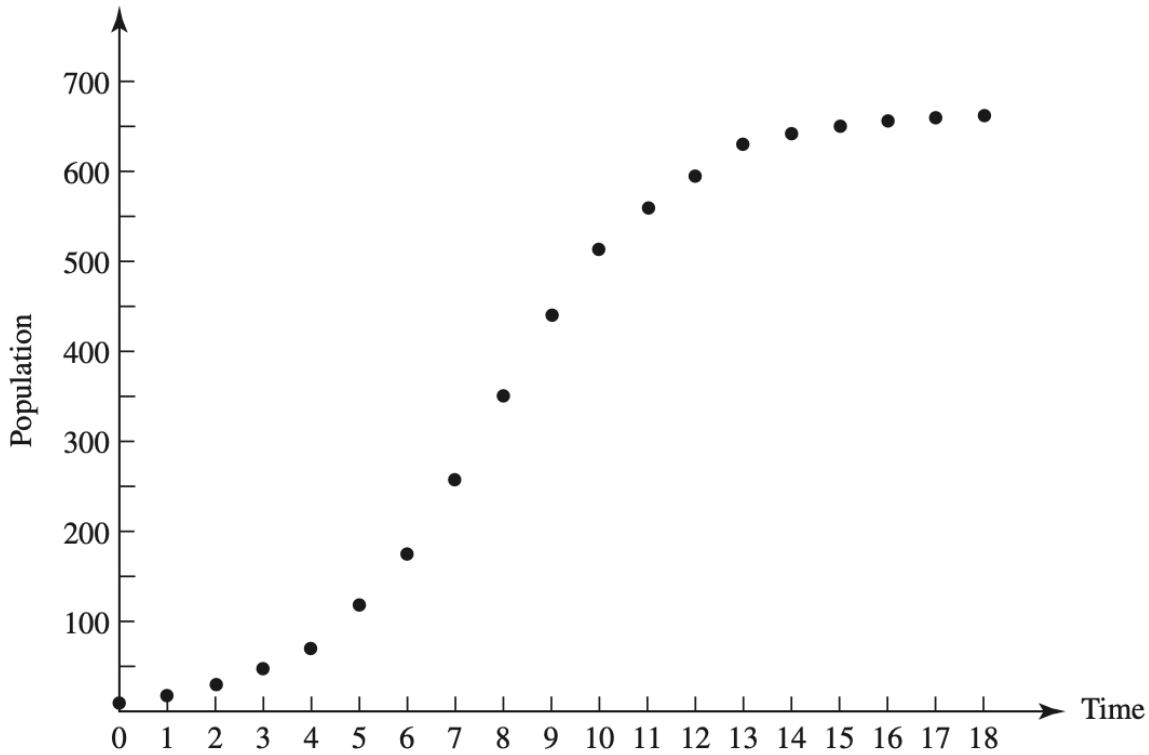
Table 4.20 A divided difference table for the data relating total vehicular stopping distance and speed

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				

Another example

Figure 4.19

A scatterplot of the "yeast growth in a culture" data



Another example (continued)

- In the second divided differences we start having negative signs
  - Data still significant in magnitude
  - The negative signs indicate a change in the concavity of the data rather than errors (see also the plot on the previous slide)
- Conclusion: the data is NOT quadratic

Table 4.22 A divided difference table for the growth of yeast in a culture

Data		Divided differences			
t_i	P_i	Δ	Δ^2	Δ^3	Δ^4
0	9.60				
		8.70			
1	18.30		1.00		
		10.70		0.92	
2	29.00		3.75		−0.31
		18.20		−0.30	
3	47.20		2.85		0.84
		23.90		3.07	
4	71.10		12.05		−1.46
		48.00		−2.77	
5	119.10		3.75		1.51
		55.50		3.28	
6	174.60		13.60		−1.51
		82.70		−2.75	
7	257.30		5.35		0.11
		93.40		−2.30	
8	350.70		−1.55		−0.05
		90.30		−2.48	
9	441.00		−9.00		0.29
		72.30		−1.32	
10	513.30		−12.95		0.94
		46.40		2.43	
11	559.70		−5.65		−0.16
		35.10		1.80	
12	594.80		−0.25		−1.40
		34.60		−3.78	
13	629.40		−11.60		1.87
		11.40		3.68	
14	640.80		−0.55		−1.10
		10.30		−0.73	
15	651.10		−2.75		0.37
		4.80		0.73	
16	655.90		−0.55		−0.20
		3.70		−0.07	
17	659.60		−0.75		
		2.20			
18	661.80				

Empirical models: conclusions

- Start by examining the data
  - Discard outliers
  - Get additional data
  - Look to see if there is a trend
- In case of a clear data trend
  - Try to build a simple, single-term model
  - Data transformations
  - Fit the model
  - Verify the model using the original data

- If one-term models fail
  - Use polynomials
  - In case of little data, use a polynomial of order $n-1$ for n data points
    - Check for oscillations especially near the end of the interval
  - In case of large data sets, use a low-order polynomial to smooth the data
    - Table of divided differences to find the degree of the polynomial
    - Fit and analyze the polynomial

Learning objectives

- Understand the concept of empirical model building (data-driven modeling)
- Obtaining data linearization through data transformation
- Calculating the Lagrangian polynomials for a given data set
- The modeling through low-order polynomials: deciding the degree and calculating the coefficients
  - deciding the degree: the technique of calculating the divided differences