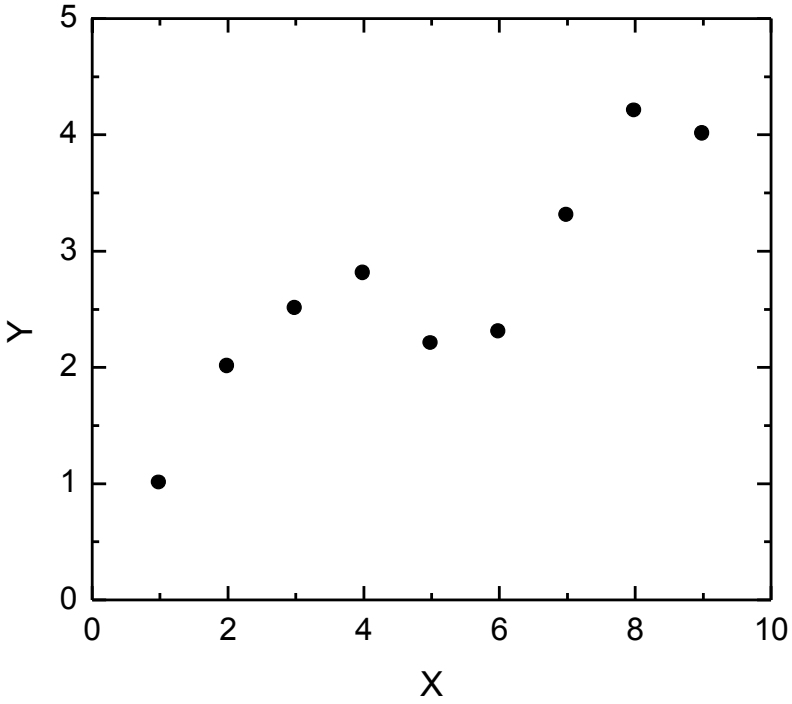
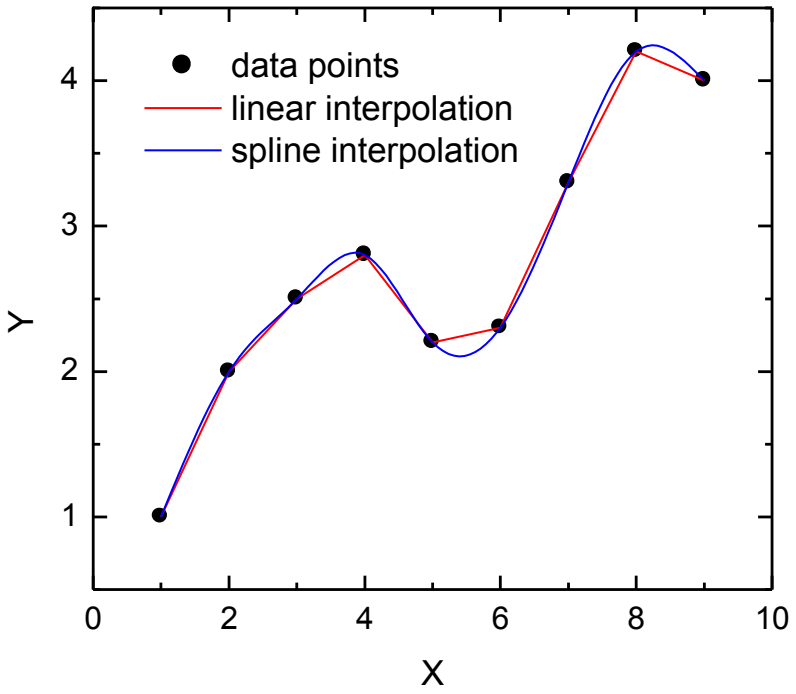
Modeling of Data

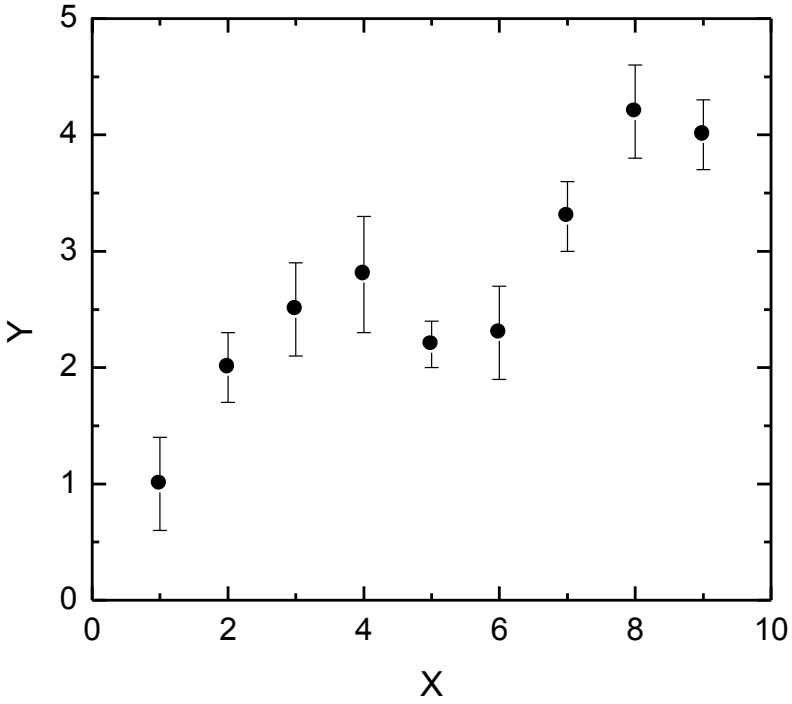


Interpolation =
local approximation

Interpolation

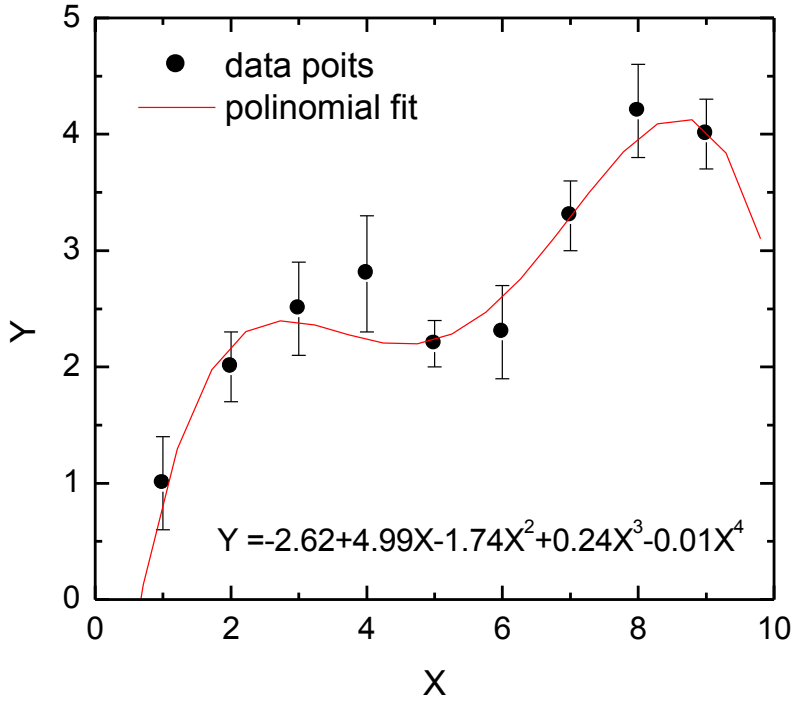


Modeling of Data

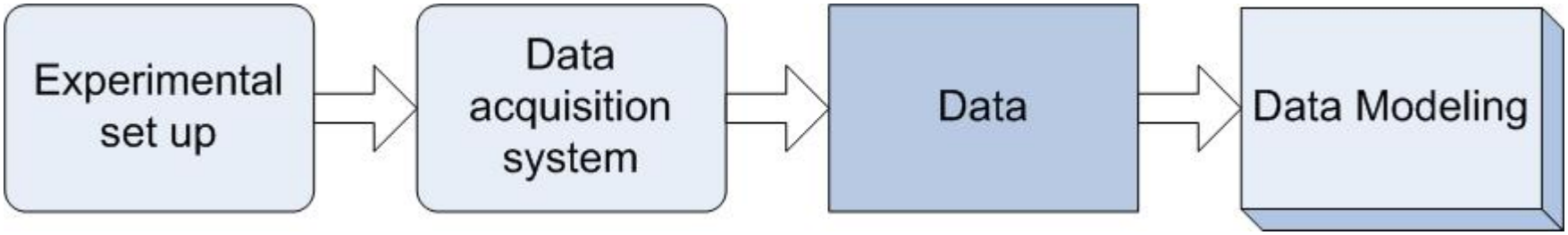


Data modeling =
global behavior

Data fit



Why data modeling?



- Condense and summarize the data
- Using data in applications
- Getting deeper insight in mechanisms

Steps in data modeling

1. Observation (experiment)
   Data are generally not exact
   (measurement errors, noise)
2. Selecting a model
   1. General: a function with adjustable parameters $y(x;a_1,a_2,\ldots a_n)$
   2. Specific: reflecting the nature of data
3. Fitting procedure

Fitting procedure should provide

- Parameters a_j in $y(x;a_1,a_2,\ldots a_m)$
- Error estimates on the parameters
- Statistical measure of goodness-of-fit

Least squares

as a maximum likehood estimator

- Suppose we are fitting N data points (x_i, y_i) $i=1, \ldots N$, to a model that has M adjustable parameters a_j, $j=1, \ldots, M$.

  $$y(x; a_1, a_2, \ldots a_M)$$

- Least-square fit
  minimize over $a_1, a_2, \ldots a_M$

$$\sum_{i=1}^{N} \left[y_i - y(x_i; a_1 \ldots a_M) \right]^2$$

Chi-Square Fitting

- If each data point (x_i, y_i) has its own, known standard deviation σ_i then the maximum likehood estimate of the model parameters is obtained by minimizing the quantity

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i; a_1 \ldots a_M)}{\sigma_i} \right)^2$$

- For linear models – closed form solutions
- For nonlinear models – *trial-and-error* procedure

Linear models

$$y(x) = c_1 f_1(x) + c_2 f_2(x) + \ldots + c_M f_M(x)$$

- Any linear model = a system of linear equations
  M – number of parameters
  N – number of data points
  N-M – degree of freedom

Fitting data to a straight line

$$y(x) = a + bx$$

$$\chi^2(a,b) = \sum_{i=1}^{N} \left(\frac{y_i - a - bx_i}{\sigma_i} \right)^2$$

- let σ_i = 1.0 (or the same for all data points)

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

Fitting data to a straight line

$$\frac{\partial \chi^2}{\partial a} = \quad 0 = \quad -2\sum_{i=1}^{N} \frac{(y_i - a - bx_i)}{\sigma_i^2}$$

$$\frac{\partial \chi^2}{\partial b} = \quad 0 = \quad -2\sum_{i=1}^{N} \frac{x_i(y_i - a - bx_i)}{\sigma_i^2}$$

- Notations

$$S \equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \qquad S_x \equiv \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \qquad S_y \equiv \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}$$

$$S_{xx} \equiv \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \qquad S_{xy} \equiv \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}$$

Fitting data to a straight line

$$
\begin{aligned}
aS &+ bS_x &&= S_y \\
aS_x &+ bS_{xx} &&= S_{xy}
\end{aligned}
$$

- the solution

$$
\begin{aligned}
a &= \frac{S_{xx}S_y - S_x S_{xy}}{\Delta} \\
b &= \frac{SS_{xy} - S_x S_{xy}}{\Delta} \\
\Delta &= SS_{xx} - (S_x)^2
\end{aligned}
$$

Probable uncertainties

- If the data are independent, then each contributes its own bit of uncertainty

$$\sigma_f^2 = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial f}{\partial y_i} \right)^2$$

$$\sigma_a^2 = S_{xx} / \Delta$$

$$\sigma_b^2 = S / \Delta$$

$$\frac{\partial a}{\partial y_i} = \frac{S_{xx} - S_x x_i}{\sigma_i^2 \Delta}$$

$$\frac{\partial b}{\partial y_i} = \frac{S_x x_i - S_x}{\sigma_i^2 \Delta}$$

correlation coefficient between uncertainty in a and b

$$r_{ab} = \frac{-S_x}{\sqrt{S S_{xx}}}$$

Goodness-of-fit

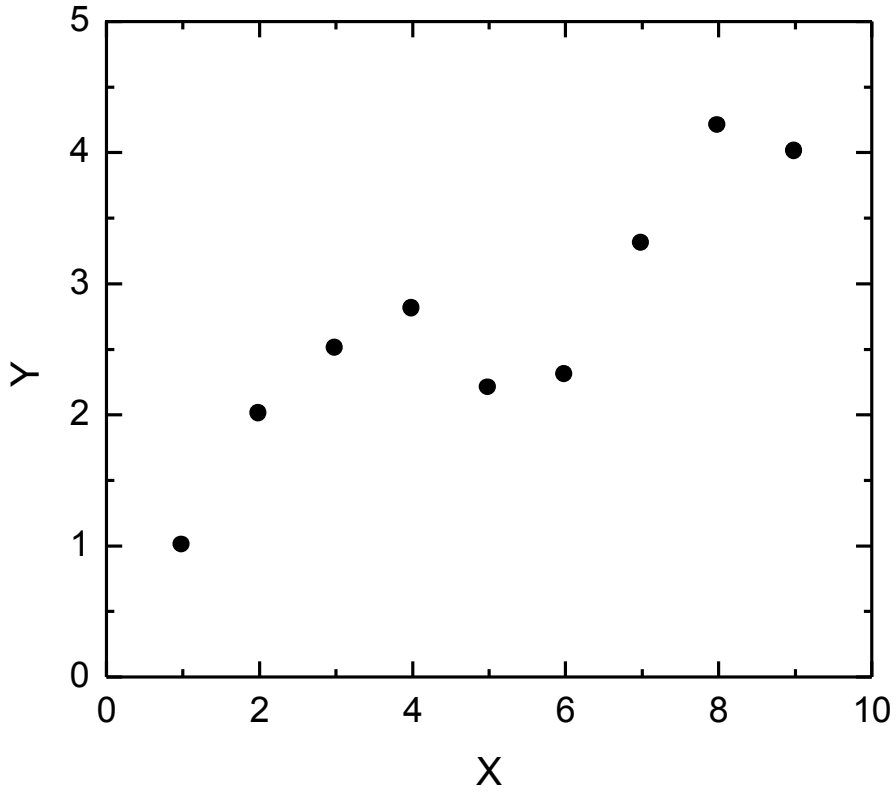
$$Q = \mathrm{gammq}\left(\frac{N-2}{2}, \frac{\chi^2}{2}\right)$$

- gammq – incomplete gamma function
- if Q>0.1 – the goodness of fit is believable
- if Q>0.001 – the fit may be acceptable
- if Q<0.001 – change the model of fitting procedure

Example: Fitting data to a straight line

Data1

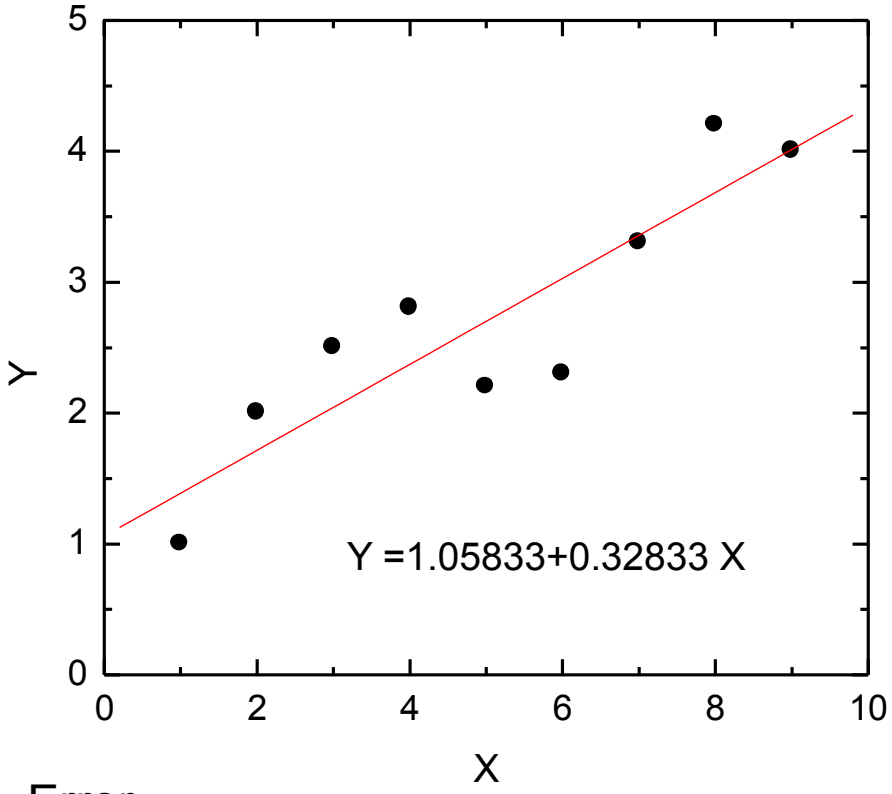
	A(X)	B(Y)
1	1	1
2	2	2
3	3	2.5
4	4	2.8
5	5	2.2
6	6	2.3
7	7	3.3
8	8	4.2
9	9	4



Example: Fitting data to a straight line

Data1

	A(X)	B(Y)
1	1	1
2	2	2
3	3	2.5
4	4	2.8
5	5	2.2
6	6	2.3
7	7	3.3
8	8	4.2
9	9	4



Parameter	Value	Error
A	1.05833	0.35504
B	0.32833	0.06309

Other issues

- Errors in both coordinates
- Multidimensional fits
- Nonlinear models (trial-and-error methods) you should guess initial values for the parameters
- Monte-Carlo simulation:

  quick-and-dirty Monte-Carlo: the bootstrap method

Programs and software

- Program libraries: minpac, lapack, slatec, sminpack, seispack, napack,…
- Excel
- Origin
- MatLab
- Systat
- Statistica