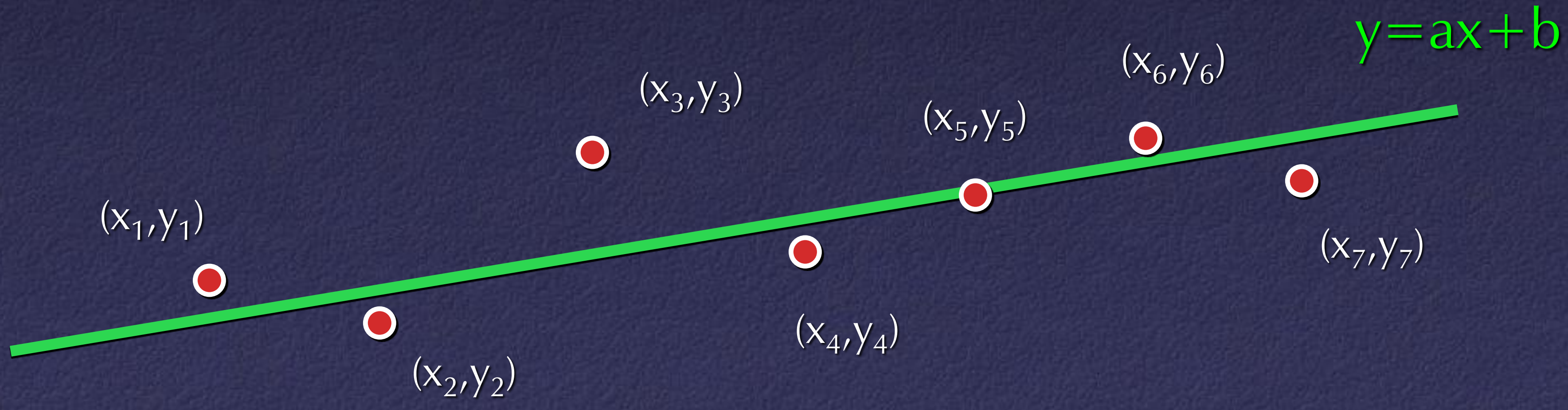
Data Modeling and Least Squares Fitting

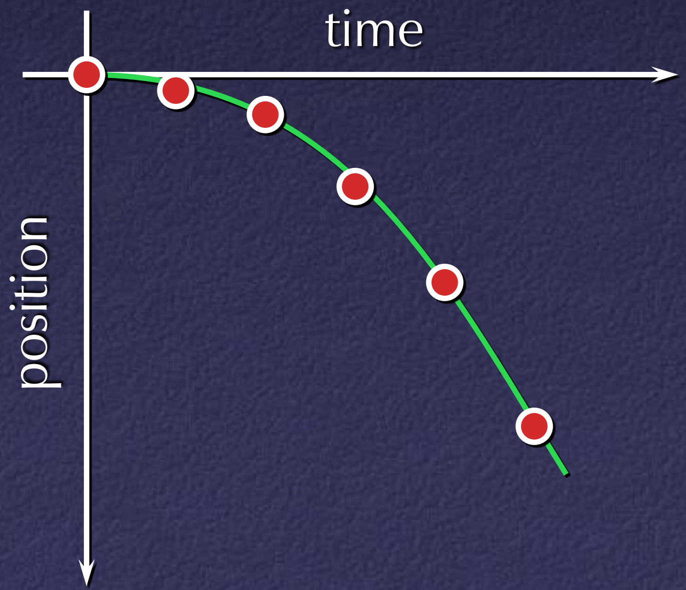
Data Modeling

- Given: data points, functional form, find constants in function
- Example: given (x_i, y_i), find line through them; i.e., find a and b in $y = ax+b$



Data Modeling

- You might do this because you actually care about those numbers…
  - Example: measure position of falling object, fit parabola

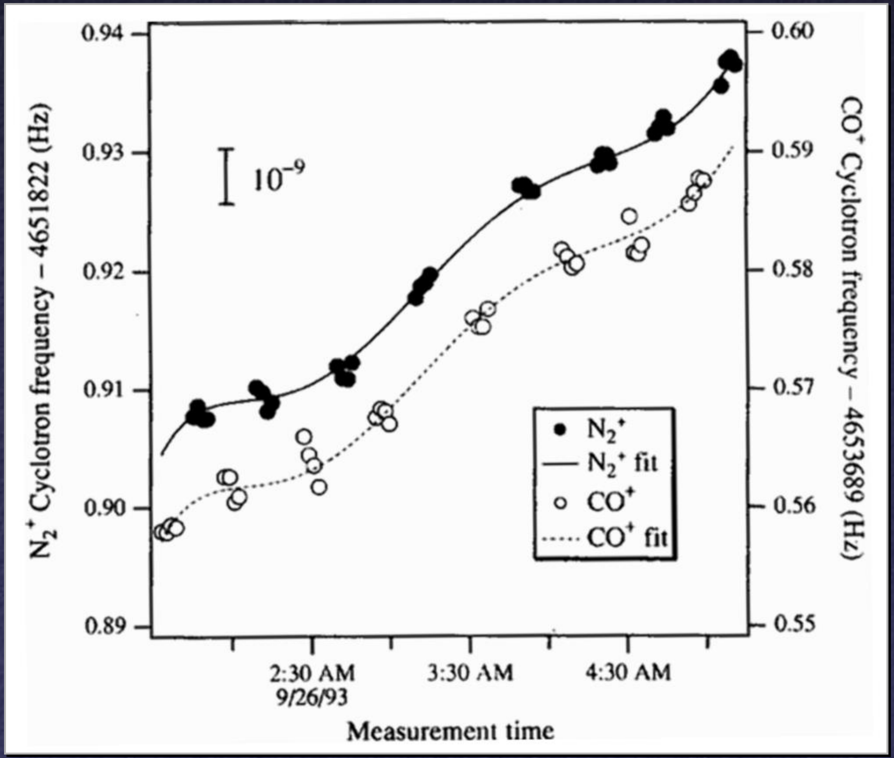


$p = -{}^{1}/_{2}\, gt^2$

$\Rightarrow$ Estimate g from fit

Data Modeling

- … or because some aspect of behavior is unknown and you want to ignore it
  - Example: measuring relative resonant frequency of two ions, want to ignore magnetic field drift



Least Squares

- Nearly universal formulation of fitting: minimize squares of differences between data and function
  - Example: for fitting a line, minimize

$$\chi^2 = \sum_i \left(y_i - (ax_i + b)\right)^2$$

    with respect to a and b
  - Most general solution technique: **take derivatives w.r.t. unknown variables, set equal to zero**

Least Squares

- Computational approaches:
  - General numerical algorithms for function minimization
  - Take partial derivatives; general numerical algorithms for root finding
  - Specialized numerical algorithms that take advantage of form of function
  - Important special case: linear least squares

Linear Least Squares

- General pattern:

$$y_i = a\,f(\vec{x}_i) + b\,g(\vec{x}_i) + c\,h(\vec{x}_i) + \cdots$$
$$\text{Given } (\vec{x}_i, y_i), \text{ solve for } a, b, c, \ldots$$

- Note that *dependence on unknowns* is linear, not necessarily function!

Solving Linear Least Squares Problem

- Take partial derivatives:

$$\chi^2 = \sum_i \left(y_i - a\,f(x_i) - b\,g(x_i) - \cdots\right)^2$$

$$\frac{\partial}{\partial a} = \sum_i -2\,f(x_i)\left(y_i - a\,f(x_i) - b\,g(x_i) - \cdots\right) = 0$$

$$a\sum_i f(x_i)\,f(x_i) + b\sum_i f(x_i)g(x_i) + \cdots = \sum_i f(x_i)\,y_i$$

$$\frac{\partial}{\partial b} = \sum_i -2g(x_i)\left(y_i - a\,f(x_i) - b\,g(x_i) - \cdots\right) = 0$$

$$a\sum_i g(x_i)\,f(x_i) + b\sum_i g(x_i)g(x_i) + \cdots = \sum_i g(x_i)\,y_i$$

Solving Linear Least Squares Problem

- For convenience, rewrite as matrix:

$$\begin{bmatrix} \sum_i f(x_i)f(x_i) & \sum_i f(x_i)g(x_i) & \cdots \\ \sum_i g(x_i)f(x_i) & \sum_i g(x_i)g(x_i) & \\ & \vdots & \end{bmatrix} \begin{bmatrix} a \\ b \\ \vdots \end{bmatrix} = \begin{bmatrix} \sum_i f(x_i)y_i \\ \sum_i g(x_i)y_i \\ \vdots \end{bmatrix}$$

- Factor:

$$\sum_i \begin{bmatrix} f(x_i) \\ g(x_i) \\ \vdots \end{bmatrix} \begin{bmatrix} f(x_i) \\ g(x_i) \\ \vdots \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} a \\ b \\ \vdots \end{bmatrix} = \sum_i y_i \begin{bmatrix} f(x_i) \\ g(x_i) \\ \vdots \end{bmatrix}$$

Linear Least Squares

- There's a different derivation of this: overconstrained linear system

$$\mathbf{A}x = b$$

$$\begin{pmatrix} \\ \\ \\ & \mathbf{A} & \\ \\ \\ \\ \end{pmatrix} \begin{pmatrix} \\ x \\ \\ \end{pmatrix} = \begin{pmatrix} \\ \\ \\ b \\ \\ \\ \\ \end{pmatrix}$$

- A has n rows and m<n columns: more equations than unknowns

Linear Least Squares

- Interpretation: find x that comes "closest" to satisfying Ax=b
  - i.e., minimize b–Ax
  - i.e., minimize |b–Ax|
  - Equivalently, minimize $|b-Ax|^2$ or (b–Ax)·(b–Ax)

$$\min \ (b - \mathbf{A}x)^{\mathrm{T}} (b - \mathbf{A}x)$$

$$\nabla\left((b - \mathbf{A}x)^{\mathrm{T}} (b - \mathbf{A}x)\right) = -2\mathbf{A}^{\mathrm{T}}(b - \mathbf{A}x) = \vec{0}$$

$$\mathbf{A}^{\mathrm{T}}\mathbf{A}x = \mathbf{A}^{\mathrm{T}}b$$

Linear Least Squares

- If fitting data to linear function:
  - Rows of A are functions of x_i
  - Entries in b are y_i
  - Minimizing sum of squared differences!

$$\mathbf{A} = \begin{bmatrix} f(x_1) & g(x_1) & \cdots \\ f(x_2) & g(x_2) & \cdots \\ & \vdots & \end{bmatrix}, \quad b = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix}$$

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} = \begin{bmatrix} \sum_i f(x_i) f(x_i) & \sum_i f(x_i) g(x_i) & \cdots \\ \sum_i g(x_i) f(x_i) & \sum_i g(x_i) g(x_i) & \cdots \\ & \vdots & \end{bmatrix}, \quad \mathbf{A}^{\mathrm{T}} b = \begin{bmatrix} \sum_i y_i f(x_i) \\ \sum_i y_i g(x_i) \\ \vdots \end{bmatrix}$$

Linear Least Squares

- Compare two expressions we've derived – equal!

$$\begin{bmatrix} \sum_i f(x_i)f(x_i) & \sum_i f(x_i)g(x_i) & \cdots \\ \sum_i g(x_i)f(x_i) & \sum_i g(x_i)g(x_i) & \cdots \\ & \vdots & \end{bmatrix} \begin{bmatrix} a \\ b \\ \vdots \end{bmatrix} = \begin{bmatrix} \sum_i y_i f(x_i) \\ \sum_i y_i g(x_i) \\ \vdots \end{bmatrix}$$

$$\sum_i \begin{bmatrix} f(x_i) \\ g(x_i) \\ \vdots \end{bmatrix} \begin{bmatrix} f(x_i) \\ g(x_i) \\ \vdots \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} a \\ b \\ \vdots \end{bmatrix} = \sum_i y_i \begin{bmatrix} f(x_i) \\ g(x_i) \\ \vdots \end{bmatrix}$$

Ways of Solving Linear Least Squares

- Option 1:

  for each x_i, y_i

  compute $f(x_i)$, $g(x_i)$, etc.

  store in row i of A

  store y_i in b

  compute $(A^T A)^{-1} A^T b$

- $(A^T A)^{-1} A^T$ is known as "pseudoinverse" of A

Ways of Solving Linear Least Squares

- Option 2:

  for each x_i, y_i

  compute $f(x_i)$, $g(x_i)$, etc.

  store in row i of A

  store y_i in b

  compute $A^T A$, $A^T b$

  solve $A^T A x = A^T b$

- These are known as the “normal equations” of the least squares problem

Ways of Solving Linear Least Squares

- These can be inefficient, since A typically much larger than A^TA and A^Tb
- Option 3:

  for each x_i, y_i

  compute $f(x_i)$, $g(x_i)$, etc.

  accumulate outer product in U

  accumulate product with y_i in v

  solve $Ux=v$

Special Case: Constant

- Let's try to model a function of the form

$$y = a$$

- In this case, $f(x_i)=1$ and we are solving

$$\sum_i [1] \; [a] = \sum_i [y_i]$$

$$\therefore \quad a = \frac{\sum_i y_i}{n}$$

- Punchline: mean is least-squares estimator for best constant fit

Special Case: Line

- Fit to y=a+bx

$$\sum_i \begin{bmatrix} 1 \\ x_i \end{bmatrix} \begin{bmatrix} 1 & x_i \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \sum_i y_i \begin{bmatrix} 1 \\ x_i \end{bmatrix}$$

$$(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1} = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix}^{-1} = \frac{\begin{bmatrix} \Sigma x_i^{\,2} & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix}}{n\Sigma x_i^{\,2} - (\Sigma x_i)^2}, \quad \mathbf{A}^{\mathrm{T}} b = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix}$$

$$a = \frac{\Sigma x_i^{\,2}\, \Sigma y_i - \Sigma x_i\, \Sigma x_i y_i}{n\Sigma x_i^{\,2} - (\Sigma x_i)^2}, \quad b = \frac{n\,\Sigma x_i y_i - \Sigma x_i\, \Sigma y_i}{n\Sigma x_i^{\,2} - (\Sigma x_i)^2}$$

x	1	3	4	7	9	12
y	0	2	5	10	12	16

	x_i	y_i	x_i^2	$x_i y_i$
	1	0	1	0
	3	2	9	6
	4	5	16	20
	7	10	49	70
	9	12	81	108
	12	16	144	192
Σ	36	45	300	396

$$A = \begin{bmatrix} 6 & 36 \\ 36 & 300 \end{bmatrix}; B = \begin{bmatrix} 45 \\ 396 \end{bmatrix}$$

$$y = \frac{2}{3}x - 3/2$$

Example: Use least-squares regression to fit a straight line to

x	1	3	5	7	10	12	13	16	18	20
y	4	5	6	5	8	7	6	9	12	11

Use least-squares regression to fit a straight line to

x	y
5	16
10	25
15	32
20	33
25	38
30	36
35	39
40	40
45	42
50	42

Fit to $y=a_0+a_1x+a_2x^2$

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{Bmatrix}$$

EXAMPLE:

Fit a second order polynomial to the following data

i	1	2	3	4	5	6
x	0	0.5	1.0	1.5	2.0	2.5
y	0	0.25	1.0	2.25	4.0	6.25

EXAMPLE: Find the least-squares parabola that fits to the following data set.

x	0	1	2	3	4	5
y	2.1	7.7	13.6	27.2	40.9	61.1

$n = 6$

$\sum x_i = 15 \qquad \sum y_i = 152.6$

$\sum x_i^2 = 55 \qquad \sum x_i y_i = 585.6$

$\sum x_i^3 = 225 \qquad \sum x_i^2 y_i = 2488.6$

$\sum x_i^4 = 979$

$a_0 = 2.479, \quad a_1 = 2.359, \quad a_2 = 1.861$

$y = 2.479 + 2.359\,x + 1.861\,x^2$

x	1	2	4	8	11	13
y	0	1	11	13	30	50

Fit to $y=ae^{bx}$

x	1,1	3,2	5,1	7,7	9,6	12,2
y	3,1	29,9	65,7	100,4	195,7	300,4