Curve fitting

MATH 3610

September 15, 2023

Curve fitting

- **Data:** Limited number of data points representing values of a function for a limited number of values of the independent variable (e.g. time, space, ...).
  - n data points (x_i, y_i) with $i = 1, \cdots, n$
- **Hypothesize the form of the function**
  - $y = f(x, p)$ where p is a $m-$vector of parameters (m parameters)
- **Curve fitting** (find the curve that has the best fit to data points) The best fit minimizes the difference between the actual value (data) and the predicted value (curve)
  - $\Rightarrow$ find the estimate of parameter values

Curve fitting: method of least-squares

Model: $f(x, p)$ with parameters p
Criterion: measure the total error in fitting a curve to data

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

- sum of squares for error = sum of squared residuals
- residual = difference between the actual value (data) and the predicted value (curve)

Aim: minimization of the sum of squared residual

$$\min_p RSS(p)$$

Result: Least-squares best fit minimizes the sum of squares of vertical distances between data points and fitting curve points.

Optimization problems

When an analytic expression of the function $\Phi(p)$ to optimize is known

Theorem

A smooth function $\Phi(p)$ attains an local minimum (resp. maximum) at $\hat{p}$ if

- the gradient $\frac{\partial \Phi(p)}{\partial p}$ vanishes at $\hat{p}$
- and the Hessian $H(p)$ with $(i,j)th$ element $\frac{\partial^2 \Phi(p)}{\partial p_i \partial p_j}$ is positive definite (resp. negative definite) at $\hat{p}$, or

$$z^T H(p) z > 0(\text{ resp. } < 0)$$

  where z is any real vector.

(If $\Phi(p)$ is non-smooth, the local extrema are at the discontinuity of $\Phi(p)$ or where the gradient $\frac{\partial \Phi(p)}{\partial p}$ is discontinuous or vanishes)

Functions of two variables

Second derivative test

To find the relative extrema of $\Phi(x, y)$

- Compute critical points (x_0, y_0) such that $\frac{\partial \Phi}{\partial x}(x_0, y_0) = 0$ and $\frac{\partial \Phi}{\partial y}(x_0, y_0) = 0$
- At the critical point (x_0, y_0):
  - If $\frac{\partial^2 \Phi}{\partial x^2}(x_0, y_0) < 0$ and $\frac{\partial^2 \Phi}{\partial x^2}(x_0, y_0)\frac{\partial^2 \Phi}{\partial y^2}(x_0, y_0) - \left(\frac{\partial^2 \Phi}{\partial x \partial y}(x_0, y_0)\right)^2 > 0$ then $\Phi(x, y)$ has a relative maximum at (x_0, y_0).
  - If $\frac{\partial^2 \Phi}{\partial x^2}(x_0, y_0) > 0$ and $\frac{\partial^2 \Phi}{\partial x^2}(x_0, y_0)\frac{\partial^2 \Phi}{\partial y^2}(x_0, y_0) - \left(\frac{\partial^2 \Phi}{\partial x \partial y}(x_0, y_0)\right)^2 > 0$ then $\Phi(x, y)$ has a relative minimum at (x_0, y_0).
  - If $\frac{\partial^2 \Phi}{\partial x^2}(x_0, y_0)\frac{\partial^2 \Phi}{\partial y^2}(x_0, y_0) - \left(\frac{\partial^2 \Phi}{\partial x \partial y}(x_0, y_0)\right)^2 < 0$ then there is a saddle point at (x_0, y_0).
  - If $\frac{\partial^2 \Phi}{\partial x^2}(x_0, y_0)\frac{\partial^2 \Phi}{\partial y^2}(x_0, y_0) - \left(\frac{\partial^2 \Phi}{\partial x \partial y}(x_0, y_0)\right)^2 = 0$, no conclusive.

Method of least-squares for models linear in parameters

Aim: find parameter values for the model which best fits data

- **Observation**: n data points (x_i, y_i) with $i = 1, \cdots, n$
- **Model**: $f(x, p)$ where p is a $m-$vector of parameters (m parameters)
- **Criterion**: sum, RSS, of squared residuals

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

- **Solution**: $\hat{p}$ such that

$$RSS(\hat{p}) = \min_p RSS(p)$$

is obtained by setting the gradient equal to zero (m parameters)

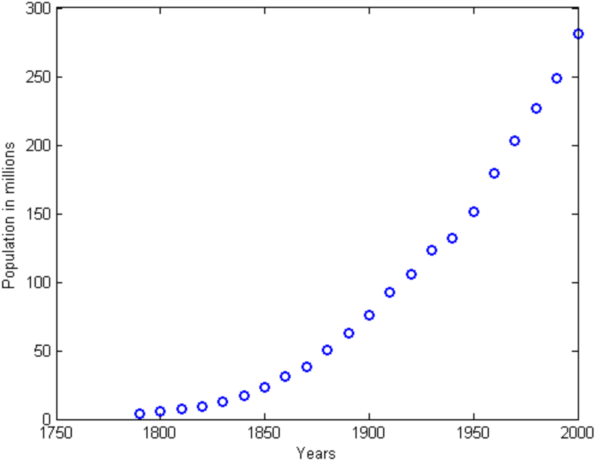
$$\frac{\partial RSS}{\partial p_j} = 0, \quad j = 1, \cdots, m$$

or

$$-2 \sum_{i=1}^{n} (y_i - f(x_i, p)) \frac{\partial f(x_i, p)}{\partial p_j} = 0, \quad j = 1, \cdots, m$$

US population (in millions) from 1790 to 2000

1790	3.929
1800	5.308
1810	7.240
1820	9.638
	12.866
	17.069
	23.192
	31.443
	38.558
	50.156
	62.948
1900	76.212
	92.228
	106.021
	123.202
	132.164
	151.325
	179.323
	203.302
	226.542
	248.709
2000	281.421



Observation: 22 data points $(x_i, y_i) =$ (year,population) with $i = 1, \dots, 22$

US population (in millions) from 1790 to 2000

Model: $f(x, p)$ where p is a $k-$vector of parameters (k parameters)

Hypothesize the form of the function f

- **Quadratic function** (x years)

$$f(x) = y = a + bx + cx^2$$

  $k = 3$ parameters to estimate a, b and c

- **Exponential function** (x years)

$$f(x) = y = a \exp^{bx}$$

  Change of variable $\ln y = Y$

$$\ln y = Y = \ln a + bx = A + bx$$

  $k = 2$ parameters to estimate A and b

Both models are linear in parameters

Nonhomogeneous linear systems

To solve a nonhomogeneous linear systems ($\det(A) \neq 0$)

$$AX = B$$

- Find the inverse of the coefficient matrix and multiply the nonhomogeneous term by the inverse:

$$X = A^{-1}B$$

- Cramer's rule

Theorem (Cramer's rule)

Consider the linear system $AX = B$ with $A \in \mathcal{M}_n$ a square matrix such that $|A| \neq 0$, $X = (x_1, \ldots, x_n)^T$ and $B = (b_1, \ldots, b_n)^T$ column vectors. Then for $i = 1, \ldots, n$,

$$x_i = \frac{|A_i|}{|A|},$$

where A_i is the matrix obtained by replacing column i in A by B.

Method to find A^{-1}

Theorem (Direct method for inversion)

Let A be a $n \times n-$matrix. If A is invertible ($|A| \neq 0$), then elementary row operations on the augmented matrix $[A|I_n]$ eventually lead to the augmented matrix $[I_n|A^{-1}]$.

Adjoint method

For $2 \times 2-$matrix: $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad A^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

US population (in millions) from 1790 to 2000

Find the minimum of

$$RSS(A, b) = \sum_{i=1}^{n} (\ln y_i - (A + bx_i))^2$$

Set the gradient of RSS to zero

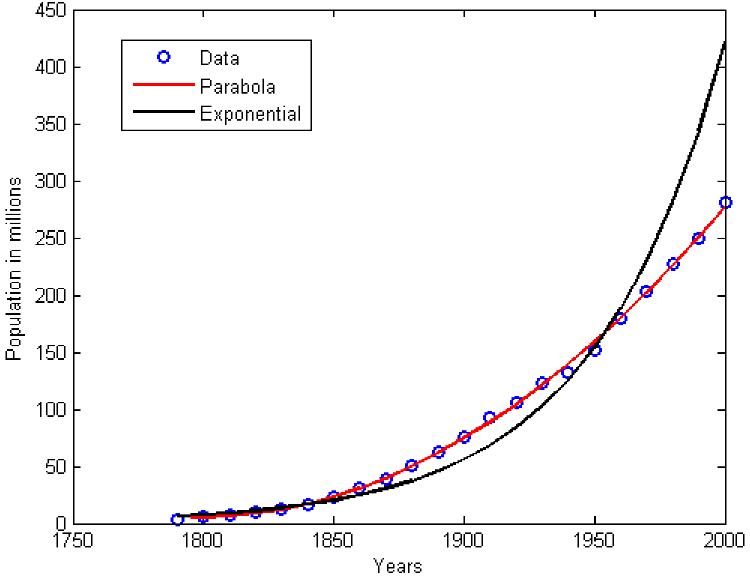
$$\sum_{i=1}^{n} (\ln y_i - (A + bx_i)) \frac{\partial (A + bx_i)}{\partial A} = \sum_{i=1}^{n} (\ln y_i - (A + bx_i)) = 0$$

$$\sum_{i=1}^{n} (\ln y_i - (A + bx_i)) \frac{\partial (A + bx_i)}{\partial b} = \sum_{i=1}^{n} (\ln y_i - (A + bx_i)) x_i = 0$$

$\hat{A}$ and $\hat{b}$ (estimate of A and b) satisfy

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} \hat{A} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \ln y_i \\ \sum_{i=1}^{n} x_i \ln y_i \end{bmatrix}$$

US population (in millions) from 1790 to 2000



Naive approach to compare models: R^2

Measure of the goodness of fit

$$R^2 = 1 - \frac{RSS/n}{\sum(y_i - \bar{y})^2/n}$$

where

- RSS = residual sum of squares
- n = sample size
- y = data

Select the model that maximizes R^2

Best fit but neglect the model complexity (select the more parameter rich model)

Only valid for linear models

To compare models: Ajusted R^2

Replacing the two variances with their unbiased estimates

Measure of the goodness of fit

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{\sum(y_i - \bar{y})^2/(n-1)}$$

where

- RSS = residual sum of squares
- n = sample size
- y = data
- p = number of parameters

Select the model that maximizes R^2_{adj}

Only valid for linear models

Estimation of parameters in mechanistic models

$$\frac{dx}{dt} = m(x, p, t),\ x(t_0) = x_0(p),\ \tilde{y} = h(x, p, t)$$

$x(t)$ vector of state variables, x_0 IC, h observable function and p vector of unknown constant parameters

To find the vector of parameter values p that minimizes the distance between measured observations and simulated observations:

- Define a distance = Scalar objective function (cost function)

$$F_{ls}(p) = \sum_{e=1}^{n_e} \sum_{o=1}^{n_o^e} \sum_{i=1}^{n_i^{e,o}} \omega_i^{e,o} (y_e^o(t_i) - \tilde{y}_o^e(t_i, p))^2$$

  n_e # of experiments, n_o^e # of observable per experiments, $n_i^{e,o}$ # of samples per observable per experiments

  $y_e^o(t_i)$ measured data, $\omega_i^{e,o}$ weights and $\tilde{y}_o^e(t_i, p)$ simulated output

- Optimization method to minimize $F_{ls}(p)$ to find $\hat{p}_{LSE}$

$$F_{ls}(\hat{p}_{LSE}) = \min_p F_{ls}(p)$$

Ordinary Differential Equations

Definition

The following is an ordinary differential equation of the first order,

$$\frac{dx}{dt} = f(t, x), \tag{E}$$

We also use the notation $x' = \frac{dx}{dt}$.

Definition

Let $J = (a, b) = \{t \in \mathbb{R} : a < t < b\}$. *A solution of the differential equation (E) on* J *is a real-valued continuously differentiable function* φ *defined on* J *such that* $(t, \varphi(t)) \in D$ *and*

$$\varphi'(t) = f(t, \varphi(t)),$$

for all $t \in J$.

(D an open connected subset of $\mathbb{R}^2$)

Initial Value Problem

Definition

Given $(\tau, \xi) \in D$, an **initial value problem** *(IVP) for (E) is given by*

$$x' = f(t,x), \quad x(\tau) = \xi \tag{I}$$

where $x(\tau) = \xi$ is the initial condition.

(A differential equation together with an initial condition form an Initial Value Problem (IVP))

Definition

A function φ is a solution of (I) if φ is a solution of the DE $x' = f(t,x)$ on some interval J containing τ and also satisfies the initial condition $\varphi(\tau) = \xi$.

Different approaches to dealing with initial value problems:

1. Analytic methods - used to obtain the exact expression of solutions of a given equation
2. Numerical methods - approximate, can be reasonably accurate. Yields approximations only locally on small intervals of the solution's domain
3. Qualitative methods - to investigate properties of solutions without necessarily finding those solutions (existence, uniqueness, stability, or chaotic or asymptotic behaviors)

Separable equations

Definition

A first order DE

$$\frac{dy}{dx} = f(x, y)$$

is said to be separable or to have separable variables if it can be expressed as follows

$$\frac{dy}{dx} = g(x)h(y).$$

(the vector field can be expressed as a product of a function of the independent variable times a function of the dependent variable)

Method to solve separable equations $\frac{dy}{dx} = g(x)h(y)$

1. Express the separable equation as follows
$$\frac{1}{h(y(x))}\frac{dy}{dx} = g(x)$$
2. As y, $\frac{dy}{dx}$, and $g(x)$ are functions of x, integrate with respect to x
$$\int \frac{1}{h(y(x))}\frac{dy}{dx}dx = \int g(x)dx$$
3. Use the Change of variable Theorem [if $u = v(x)$, $\int f(v(x))v'(x)dx = \int f(u)du$] for the left side with $u = y(x)$
$$\int \frac{1}{h(u)}du = \int g(x)dx$$
$$\int \frac{1}{h(y)}dy = \int g(x)dx$$
4. Integrate
$$H(y) = G(x) + c \tag{1}$$
c is the combination of the left and right integration constants, H and G are antiderivatives of $\frac{1}{h(y)}$ and $g(x)$ respectively.

Identifiability

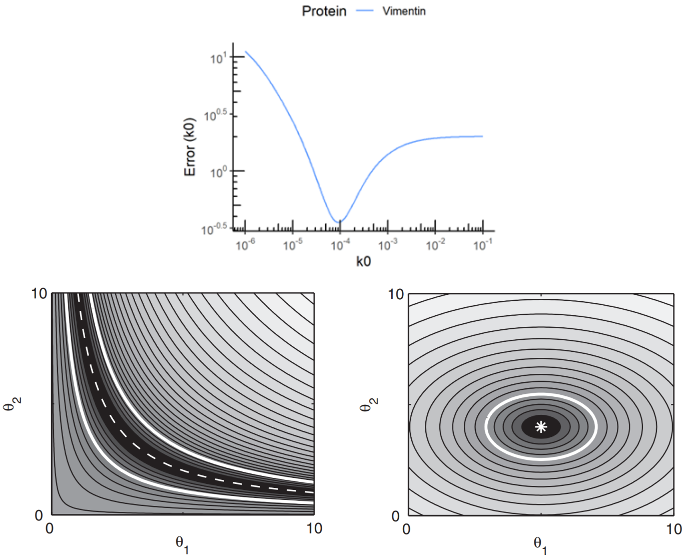
Can unknown model parameters uniquely be determined by parameter estimation from measured data? $\Rightarrow$ Identifiability

Two problems:

1. the larger the number of unknown parameters in a model, the larger the amount of quantitative data necessary to determine meaningful values for these parameters (Practical identifiability)
2. even if appropriate experimental data are available, model parameters may not be uniquely identifiable (Structural identifiability)

Identifiability

Profiles of error as a function of parameters to be estimated



Optimization methods

When an analytic expression of the function to optimize is unknown

- Local optimization methods:
  - gradient descent-based methods: Levenberg-Marquardt or Gauss-Newton
  - derivative-free local search methods: Nelder-Mead method
  - only find a global optimum for appropriate starting points
  - converge to local optima
  - suboptimal solutions
- Global optimization methods:
  - simulated annealing
  - genetic algorithm
  - particle swarm

Pitt and Banga (2019) BMC Bioinformatics. 20:82. Sagar *et al.* (2018) BMC Systems Biology 12:87.

Outline

The least squares problem (simplest version)

Definition

Given a collection of points $(x_1, y_1), \ldots, (x_n, y_n)$, find the coefficients a, b of the line $y = a + bx$ such that

$$\|\mathbf{e}\| = \sqrt{\varepsilon_1^2 + \cdots + \varepsilon_n^2} = \sqrt{(y_1 - \tilde{y}_1)^2 + \cdots + (y_n - \tilde{y}_n)^2}$$

is minimal, where $\tilde{y}_i = a + bx_i$ for $i = 1, \ldots, n$

We can solve this by brute force using, e.g., a genetic algorith to minimise $\|e\|$. Let us now see how to solve this problem "properly"

For a data point $i = 1, \ldots, n$

$$\varepsilon_i = y_i - \tilde{y}_i = y_i - (a + bx_i)$$

So if we write this for all data points,

$$\begin{aligned} \varepsilon_1 &= y_1 - (a + bx_1) \\ &\vdots \\ \varepsilon_n &= y_n - (a + bx_n) \end{aligned}$$

In matrix form

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

with

$$\mathbf{e} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, A = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \mathbf{x} = \begin{pmatrix} a \\ b \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

The least squares problem (reformulated)

Definition (Least squares solutions)

Consider a collection of points $(x_1, y_1), \ldots, (x_n, y_n)$, a matrix $A \in \mathcal{M}_{mn}$, $\mathbf{b} \in \mathbb{R}^m$. A **least squares solution** of $A\mathbf{x} = \mathbf{b}$ is a vector $\tilde{\mathbf{x}} \in \mathbb{R}^n$ s.t.

$$\forall \mathbf{x} \in \mathbb{R}^n, \quad \|\mathbf{b} - A\tilde{\mathbf{x}}\| \leq \|\mathbf{b} - A\mathbf{x}\|$$

Needed to solve the problem

Definition (Best approximation)

Let V be a vector space, $W \subset V$ and $\mathbf{v} \in V$. The **best approximation** to $\mathbf{v}$ in W is $\tilde{\mathbf{v}} \in W$ s.t.

$$\forall \mathbf{w} \in W, \mathbf{w} \neq \tilde{\mathbf{v}}, \quad \|\mathbf{v} - \tilde{\mathbf{v}}\| < \|\mathbf{v} - \mathbf{w}\|$$

Theorem (Best approximation theorem)

Let V be a vector space with an inner product, $W \subset V$ and $\mathbf{v} \in V$. Then $\mathrm{proj}_W(\mathbf{v})$ *is the best approximation to* $\mathbf{v}$ *in* W

Let us find the least squares solution

$\forall \mathbf{x} \mathbb{R}^n$, $A\mathbf{x}$ is a vector in the **column space** of A (the space spanned by the vectors making up the columns of A)

Since $\mathbf{x} \in \mathbb{R}^n$, $A\mathbf{x} \in \text{col}(A)$

$\implies$ least squares solution of $A\mathbf{x} = \mathbf{b}$ is a vector $\tilde{\mathbf{y}} \in \text{col}(A)$ s.t.

$$\forall \mathbf{y} \in \text{col}(A), \quad \|\mathbf{b} - \tilde{\mathbf{y}}\| \leq \|\mathbf{b} - \mathbf{y}\|$$

This looks very much like Best approximation and Best approximation theorem

Putting things together

We just stated: The least squares solution of $A\mathbf{x} = \mathbf{b}$ is a vector $\tilde{\mathbf{y}} \in \mathrm{col}(A)$ s.t.

$$\forall \mathbf{y} \in \mathrm{col}(A), \quad \|\mathbf{b} - \tilde{\mathbf{y}}\| \leq \|\mathbf{b} - \mathbf{y}\|$$

We know (reformulating a tad):

Theorem (Best approximation theorem)

Let V be a vector space with an inner product, $W \subset V$ and $\mathbf{v} \in V$. Then $\mathrm{proj}_W(\mathbf{v}) \in W$ is the best approximation to $\mathbf{v}$ in W, i.e.,

$$\forall \mathbf{w} \in W, \mathbf{w} \neq \mathrm{proj}_W(\mathbf{v}), \quad \|\mathbf{v} - \mathrm{proj}_W(\mathbf{v})\| < \|\mathbf{v} - \mathbf{w}\|$$

$\implies W = \mathrm{col}(A)$, $\mathbf{v} = \mathbf{b}$ and $\tilde{\mathbf{y}} = \mathrm{proj}_{\mathrm{col}(A)}(\mathbf{b})$

So if $\tilde{\mathbf{x}}$ is a least squares solution of $A\mathbf{x} = \mathbf{b}$, then

$$\tilde{\mathbf{y}} = A\tilde{\mathbf{x}} = \mathrm{proj}_{\mathrm{col}(A)}(\mathbf{b})$$

We have

$$\mathbf{b} - A\tilde{\mathbf{x}} = \mathbf{b} - \mathrm{proj}_{\mathrm{col}(A)}(\mathbf{b}) = \mathrm{perp}_{\mathrm{col}(A)}(\mathbf{b})$$

and it is easy to show that

$$\mathrm{perp}_{\mathrm{col}(A)}(\mathbf{b}) \perp \mathrm{col}(A)$$

So for all columns $\mathbf{a}_i$ of A

$$\mathbf{a}_i \cdot (\mathbf{b} - A\tilde{\mathbf{x}}) = 0$$

which we can also write as $\mathbf{a}_i^T(\mathbf{b} - A\tilde{\mathbf{x}}) = 0$

For all columns $\mathbf{a}_i$ of A,

$$\mathbf{a}_i^T(\mathbf{b} - A\tilde{\mathbf{x}}) = 0$$

This is equivalent to saying that

$$A^T(\mathbf{b} - A\tilde{\mathbf{x}}) = \mathbf{0}$$

We have

$$\begin{aligned} A^T(\mathbf{b} - A\tilde{\mathbf{x}}) = \mathbf{0} &\iff A^T\mathbf{b} - A^TA\tilde{\mathbf{x}} = \mathbf{0} \\ &\iff A^T\mathbf{b} = A^TA\tilde{\mathbf{x}} \\ &\iff A^TA\tilde{\mathbf{x}} = A^T\mathbf{b} \end{aligned}$$

The latter system constitutes the **normal equations** for $\tilde{\mathbf{x}}$

Least squares theorem

Theorem (Least squares theorem)

$A \in \mathcal{M}_{mn}$, $\mathbf{b} \in \mathbb{R}^m$. *Then*

1. $A\mathbf{x} = \mathbf{b}$ *always has at least one least squares solution* $\tilde{\mathbf{x}}$
2. $\tilde{\mathbf{x}}$ *least squares solution to* $A\mathbf{x} = \mathbf{b} \iff \tilde{\mathbf{x}}$ *is a solution to the normal equations* $A^T A\tilde{\mathbf{x}} = A^T\mathbf{b}$
3. A *has linearly independent columns* $\iff A^T A$ *invertible.*
   In this case, the least squares solution is unique and

$$\tilde{\mathbf{x}} = \left(A^T A\right)^{-1} A^T\mathbf{b}$$

We have seen 1 and 2, we will not show 3 (it is not hard)

Suppose we want to fit something a bit more complicated..

For instance, instead of the affine function

$$y = a + bx$$

suppose we want to do the quadratic

$$y = a_0 + a_1 x + a_2 x^2$$

or even

$$y = k_0 e^{k_1 x}$$

How do we proceed?

Fitting the quadratic

We have the data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and want to fit

$$y = a_0 + a_1 x + a_2 x^2$$

At (x_1, y_1),

$$\tilde{y}_1 = a_0 + a_1 x_1 + a_2 x_1^2$$

$\vdots$

At (x_n, y_n),

$$\tilde{y}_n = a_0 + a_1 x_n + a_2 x_n^2$$

In terms of the error

$$
\begin{aligned}
\varepsilon_1 &= y_1 - \tilde{y}_1 = y_1 - (a_0 + a_1 x_1 + a_2 x_1^2) \\
&\vdots \\
\varepsilon_n &= y_n - \tilde{y}_n = y_n - (a_0 + a_1 x_n + a_2 x_n^2)
\end{aligned}
$$

i.e.,

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

where

$$
\mathbf{e} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, A = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \mathbf{x} = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}
$$

Theorem 8 applies, with here $A \in \mathcal{M}_{n3}$ and $\mathbf{b} \in \mathbb{R}^n$

Fitting the exponential

Things are a bit more complicated here

If we proceed as before, we get the system

$$y_1 = k_0 e^{k_1 x_1}$$
$$\vdots$$
$$y_n = k_0 e^{k_1 x_n}$$

$e^{k_1 x_i}$ is a nonlinear term, it cannot be put in a matrix

However: take the ln of both sides of the equation

$$\ln(y_i) = \ln(k_0 e^{k_1 x_i}) = \ln(k_0) + \ln(e^{k_1 x_i}) = \ln(k_0) + k_1 x_i$$

If $y_i, k_0 > 0$, then their ln are defined and we're in business..

$$\ln(y_i) = \ln(k_0) + k_1 x_i$$

So the system is

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

with

$$A = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \mathbf{x} = \begin{pmatrix} k_1 \end{pmatrix}, \mathbf{b} = \begin{pmatrix} \ln(k_0) \end{pmatrix} \text{ and } \mathbf{y} = \begin{pmatrix} \ln(y_1) \\ \vdots \\ \ln(y_n) \end{pmatrix}$$