4.2 A Survey of Models

The "art" of data modeling includes a wide variety of models – not just linear models.

Ex. Periodic Data should, naturally, be fit with a periodic model.

We fit the following periodic data set using a linear combination of sine & cosine functions.

Time of day	T	Temp (°C)
12 mid.	0	-2.2
3 am	1/8	-2.8
6 am	1/4	-6.1
9 am	3/8	-3.9
12 pm	1/2	0.0
3 pm	5/8	1.1
6 pm	3/4	-.6
9 pm	7/8	-1.1

$\beta = \frac{2\pi}{\text{Period}}$ (P=1) 24 hours

We choose the model:

$$y = c_1 + c_2 \cos\left(\frac{2\pi t}{(\beta)}\right) + c_3 \sin\left(\frac{2\pi t}{(\beta)}\right)$$

We substitute the data into the model: (Note the system is overdetermined here, since $n > p$).

$$c_1 + c_2 \cos(2\pi \cdot 0) + c_3 \sin(2\pi \cdot 0) = -2.2$$
$$c_1 + c_2 \cos(2\pi \cdot \tfrac{1}{8}) + c_3 \sin(2\pi \cdot \tfrac{1}{8}) = -2.8$$
$$c_1 + c_2 \cos(2\pi \cdot \tfrac{1}{4}) + c_3 \sin(2\pi \cdot \tfrac{1}{4}) = -6.1$$
$$c_1 + c_2 \cos(2\pi \cdot \tfrac{3}{8}) + c_3 \sin(2\pi \cdot \tfrac{3}{8}) = -3.9$$
$$c_1 + c_2 \cos(2\pi \cdot \tfrac{1}{2}) + c_3 \sin(2\pi \cdot \tfrac{1}{2}) = 0.0$$

$$c_1 + c_2 \cos\left(2\pi \cdot \frac{5}{8}\right) + c_3 \sin\left(2\pi \cdot \frac{5}{8}\right) = 1.1$$

$$c_1 + c_2 \cos\left(2\pi \cdot \frac{3}{4}\right) + c_3 \sin\left(2\pi \cdot \frac{3}{4}\right) = -0.6$$

$$c_1 + c_2 \cos\left(2\pi \cdot \frac{7}{8}\right) + c_3 \sin\left(2\pi \cdot \frac{7}{8}\right) = -1.1$$

$$\to A = \begin{bmatrix} 1 & \cos(0) & \sin(0) \\ 1 & \cos\left(\frac{\pi}{4}\right) & \sin\left(\frac{\pi}{4}\right) \\ & \vdots & \\ 1 & \cos\left(\frac{7\pi}{4}\right) & \sin\left(\frac{7\pi}{4}\right) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & \sqrt{2}/2 & \sqrt{2}/2 \\ 1 & 0 & 1 \\ & \vdots & \\ 1 & \sqrt{2}/2 & -\sqrt{2}/2 \end{bmatrix} \quad \& \quad \vec{b} = \begin{bmatrix} -2.2 \\ -2.8 \\ \vdots \\ -.6 \\ -1.1 \end{bmatrix}$$

The normal equations: $A^T A c = A^T b$

$$\to \begin{bmatrix} 8 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} -15.6 \\ -2.9778 \\ -10.2376 \end{bmatrix} \to \begin{matrix} c_1 = -1.95 \\ c_2 = -.7445 \\ c_3 = -2.5594 \end{matrix}$$

$$\to \hat{y} = -1.95 - 0.7445 \cos 2\pi t - 2.5594 \sin 2\pi t$$

RMSE $\approx$ 1.063

Ex. 2 Now fit the same data to the improved model:

$$\hat{y} = c_1 + c_2 \cos(2\pi t) + c_3 \sin(2\pi t) + c_4 \cos(4\pi t)$$

The system of equations for this model is as follows:

8 eq's

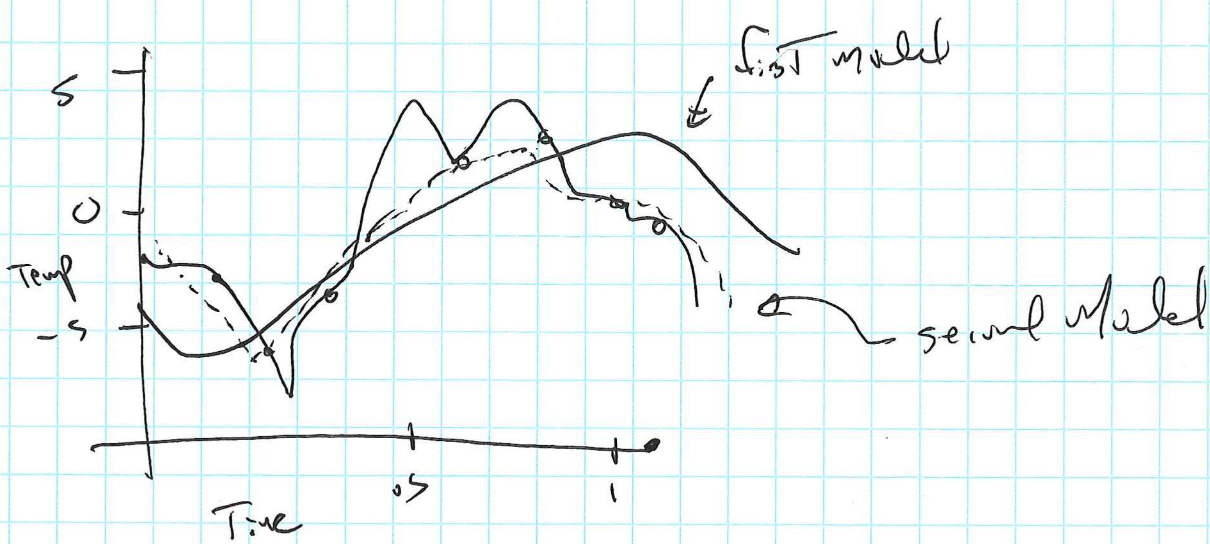
$$\begin{cases} c_1 + c_2 \cos 2\pi(0) + c_3 \sin 2\pi(0) + c_4 \cos 4\pi(0) = -2.2 \\ \vdots \\ c_1 + c_2 \cos 2\pi\left(\frac{7}{8}\right) + c_3 \sin\left(2\pi \cdot \frac{7}{8}\right) + c_4 \cos 4\pi\left(\frac{7}{8}\right) = -1.1 \end{cases}$$

The corresponding Normal Equations are:

$$\underbrace{\begin{bmatrix} 8 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}}_{A^T A} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \underbrace{\begin{bmatrix} -15.6 \\ -2.9778 \\ -10.2376 \\ 4.5 \end{bmatrix}}_{A^T b}$$

$$\rightarrow \quad y = -1.95 - 0.7445 \cos 2\pi t - 2.5594 \sin 2\pi t + 1.125 \cos 4\pi t$$

RMSE ≅ 0.705 → less than previous model!



Data Linearization

Exponential growth of a population is implied when:

$$\frac{dp}{dt} = kp(t)$$

i.e. the growth rate is proportional to population size.

Under "perfect conditions", when the environment remains unchanged & the population size is well below the environmental carrying capacity, we have the uninhibited growth model:

$$y = c_1 e^{c_2 t}$$

Note that this model cannot be directly fit by least squares because c_2 does not appear linearly in the model equation. (So we can't write the system as $A\vec{x} = \vec{b}$).

There are ② remedies here: ① Directly minimize the least squares error (Gauss-Newton §4.5); ② "Linearize" the model — which we do now.

$$y = c_1 e^{c_2 t} \rightarrow \ln y = \ln(c_1 e^{c_2 t}) = \ln(c_1) + \ln(e^{c_2 t})$$

$$= \ln(c_1) + c_2 t$$

Now make variable substitution: $k = \ln(c_1)$

$$= k + c_2 t \rightarrow \text{Model is linear in } k \text{ \& } c_2!$$

In summary: $\ln y = \underbrace{k + c_2 T}_{\text{"Linear"}}$

Next: (1) Solve the corresponding normal equations for k & c_2

(2) Revert back to c_1 → i.e. set $c_1 = e^k$

A few notes on Linearization:

Our solution here involved changing the original problem.

Originally, we would ~~[illegible]~~ minimize:

$$(c_1 e^{c_2 T_1} - y_1)^2 + \dots + (c_1 e^{c_2 T_m} - y_m)^2, \quad (1)$$

i.e. the sum of squares of the residuals for the model: $y = c_1 e^{c_2 T}$.

Here, conversely, we solve the revised problem minimizing the least squares error in "log space" — i.e. we minimize:

$$(\ln c_1 + c_2 T_1 - \ln y_1)^2 + \dots + (\ln c_1 + c_2 T_m - \ln y_m)^2 \quad (2)$$

Observe that there are two different minimization problems, with different solutions!

Q: Which is the "better" method? It depends — it may be more natural, depending on the problem, to evaluate the fitness of the model after moving to log space.

Ex. Use model linearization to find the best least squares exponential fit: $y = c_1 e^{c_2 T}$ for the data:

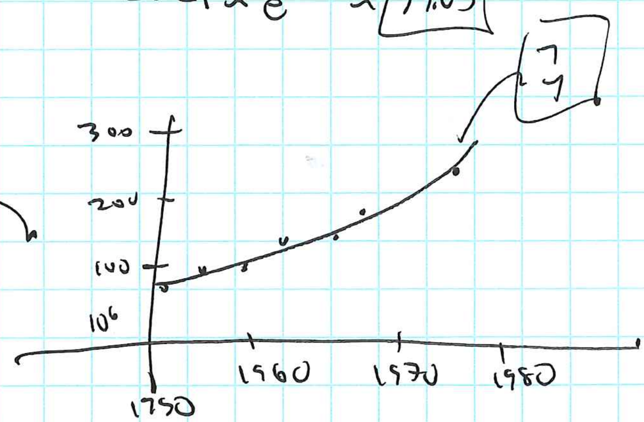
year	world automobile pop. $(\cdot 10^6)$
(T=0) 1950	53.05
1955	73.04
1960	98.31
1965	139.78
1970	193.48
1975	260.20
1980	320.39

→ solving the linear least squares problem: $k_1 \approx 3.9896$, $c_2 \approx .06158$

$\rightarrow c_1 \approx e^{3.98} \approx 54.03$

$$\rightarrow y = 54.03 e^{.06152T}$$

RMSE ≈ 9.56



Ex. Data set is the number of ~~transistors~~ in Intel CPU since the 1970s.

Use the exponential model: $y = c_1 e^{c_2 T}$

Moore's Law ← comp. power doubles every 2 years

CPU	year	# Transistors
4004	1971	2,250
8008	'72	2,500
8080	'74	5,000
8086	'78	29,000
286	'82	120k
386	'85	275k
486	'89	1.18M
Pent	'93	3.1M
Pent II	'97	7.5M
Pent III	'99	24M
Pent IV	2000	42M
Itanium	2002	220M
Itanium II	2003	410M

$\ln y = k + c_2 T$ $\quad (T = 0 \leftrightarrow 1970)$

$$k + c_2(1) = \ln 2250$$
$$\vdots$$
$$k + c_2(8) = \ln 29k$$
$$\vdots$$

$$\rightarrow A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 33 \end{bmatrix}, \quad \vec{b} = \begin{bmatrix} \ln 2250 \\ \vdots \\ \ln 41M \end{bmatrix}$$

$$A^T A \vec{x} = A^T \vec{b} \rightarrow \begin{bmatrix} 13 & 235 \\ 235 & 5927 \end{bmatrix} \begin{bmatrix} k \\ c_2 \end{bmatrix} = \begin{bmatrix} 176 \\ 3793.23 \end{bmatrix}$$

→ solving: $k \approx 7.197$, $c_2 \approx 0.3546$ → $c_1 = e^k \approx 1335.3$

$$\hat{y} = 1335.3e^{0.3546T}$$

Doubling Time $\approx \frac{\ln 2}{c_2} \approx 1.95$ years

Another important example with non-linear coefficients is the power law model:

$$y = c_1 T^{c_2}$$

We "linearize" the model: $\ln y = \ln(c_1 T^{c_2}) = \ln(c_1) + \ln(T^{c_2})$

$$= \ln(c_1) + c_2 \ln(T)$$

$$= k + c_2 \ln(T)$$ → "linear" WRT k, c_2.

This gives the following linear system:

$$\left.\begin{aligned} k + c_2\ln(T_1) &= \ln(y_1) \\ k + c_2\ln(T_2) &= \ln(y_2) \\ &\vdots \\ k + c_2\ln(T_n) &= \ln(T_n) \end{aligned}\right\} \xrightarrow{\text{Matrix Form}} A = \begin{bmatrix} 1 & \ln(T_1) \\ 1 & \ln(T_2) \\ \vdots & \vdots \\ 1 & \ln(T_n) \end{bmatrix} \;\&\; \vec{b} = \begin{bmatrix} \ln y_1 \\ \vdots \\ \ln y_n \end{bmatrix}$$

Ex. Use linearization to fit the given height/weight data with a power law model.

$$\hat{W} = 16.3 H^{2.42}$$

age (years)	height (m)	weight (kg)
2	.9120	13.7
3	.98	15.9
4	1.06	18.5
5	1.13	
6	1.19	21.3
7	1.26	23.5
8	1.32	32.7
9	1.38	36.0
10	1.41	38.6
11	1.49	43.7

·) The time course of <u>drug concentration</u> y in the bloodstream is well-defined by:

$$y = c_1 T e^{c_2 T}$$

, T: <u>Time</u> after administering drug

(<u>Half-life</u>) Time when $y = \frac{1}{2} y$

<u>Linearize</u>: →

$$\ln y = \ln(c_1) + \ln(T) + \ln\left(e^{c_2 T}\right)$$

$$\rightarrow \ln y = \ln(c_1) + \ln(T) + c_2 T$$

$$\rightarrow \boxed{k + c_2 T = \ln y - \ln T} \qquad (k = \ln(c_1)).$$

linear wrt parameters: k, c_2

<u>Matrix Equation</u>

$$A = \begin{bmatrix} 1 & T_1 \\ \vdots & \vdots \\ 1 & T_m \end{bmatrix} \;\&\; \vec{b} = \begin{bmatrix} \ln y_1 - \ln T_1 \\ \vdots \\ \ln y_m - \ln T_m \end{bmatrix}$$

Ex. Fit the model with the measured level of the drug norfluoxetine in a patient's bloodstream.

hour	Concentration (ng/ml)
1	8.0
2	12.3
3	15.5
4	16.8
5	17.1
6	15.8
7	15.2
8	14.0

→ $k \approx 2.28 \rightarrow c_1 \approx e^{2.28} \approx \underline{9.77}$

$c_2 \approx -.215$

$$\rightarrow \boxed{\hat{y} = 9.77\, T\, e^{-.215 T}}$$