Least Squares Regression

Data given :- $y_i := y(x_i)$ are given for all x_i ; $i = 1, 2, \dots, n$

Goal :- We want to find the curve of best fit of the form $y = a + bf(x) + cg(x)$ that most suitably describes the data (x_i, y_i). Here a, b, c are constants and $f(x)$ and $g(x)$ are model f^ns. of our choice

Plan :- Unleash the method of least sqs. to minimize the objective f^n

$$e = r^2 = \sum_{i=1}^{n} \{ y_i - (a + bf(x_i) + cg(x_i)) \}^2$$

from calculus.

$$\frac{\partial e}{\partial a} = \frac{\partial e}{\partial b} = \frac{\partial e}{\partial c} = 0 \text{ to find our optimal } a, b, c$$

$$\frac{\partial e}{\partial a} = \sum_{i=1}^{n} 2\{ y_i - (a + bf(x_i) + cg(x_i)) \}(-1) = 0$$

$$\left.\begin{aligned} &\Rightarrow \sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} 1 + b\sum_{i=1}^{n} f(x_i) + c\sum_{i=1}^{n} g(x_i) \\ &\Rightarrow \sum_i y_i = a\sum_{i=1}^{n} 1 + b\sum_i f_i + c\sum_i g_i \end{aligned}\right\} \text{— ①}$$

$$\frac{\partial e}{\partial b} = \sum_{i=1}^{n} 2\{y_i - (a + bf_i + cg_i)\}(-f_i) = 0 .$$

$$\Rightarrow \sum_i y_i f_i = a\sum_i f_i + b\sum_i f_i^2 + c\sum_i f_i g_i \quad \text{— ②}$$

and,

$$\frac{\partial e}{\partial c} = \sum_{i=1}^{n} 2\{y_i - (a + bf_i + cg_i)\}(-g_i) = 0$$

$$\Rightarrow \sum_i y_i g_i = a\sum_i g_i + b\sum_i f_i g_i + c\sum_i g_i^2 \quad \text{— ③}$$

Eqn ①, ② & ③ can be written in matrix form as

$$\underbrace{\begin{pmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} f_i & \sum_{i=1}^{n} g_i \\ \sum_{i=1}^{n} f_i & \sum_{i=1}^{n} f_i^2 & \sum_{i=1}^{n} f_i g_i \\ \sum_{i=1}^{n} g_i & \sum_{i=1}^{n} f_i g_i & \sum_{i=1}^{n} g_i^2 \end{pmatrix}}_{\text{Call this } \Lambda} \underbrace{\begin{pmatrix} a \\ b \\ c \end{pmatrix}}_{\alpha} = \underbrace{\begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i f_i \\ \sum_{i=1}^{n} y_i g_i \end{pmatrix}}_{\chi}$$

So solution is $\boxed{\alpha = \Lambda^{-1} \chi}$

eg. Consider the following data

Hrs. of sunshine x_i	No. of ice-creams sold y_i
2	4
3	5
5	7
7	10
9	15

Here $n = 5$

1) Fit a line of best fit.
2) Estimate based on the line of best fit, how many ice-creams will be sold in a day w/ 8 hrs of sunshine.

Soln:- $y = a + bx$ is the line of best fit ; so $f(x) = x$, $g(x) = 0$.

$$\begin{pmatrix} \sum_{i=1}^{5} 1 & \sum_{i=1}^{5} x_i \\ \sum_{i=1}^{5} x_i & \sum_{i=1}^{5} x_i^2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{5} y_i \\ \sum_{i=1}^{5} y_i x_i \end{pmatrix} \Rightarrow \Lambda = \begin{pmatrix} 5 & 26 \\ 26 & 168 \end{pmatrix} ; \chi = \begin{pmatrix} 41 \\ 263 \end{pmatrix}$$

$$\text{so } \Lambda^{-1} = \begin{pmatrix} 1.0244 & -0.1585 \\ -0.1585 & 0.0305 \end{pmatrix}$$

so $\alpha = \begin{pmatrix} a \\ b \end{pmatrix} = \Lambda^{-1} X = \begin{pmatrix} 0.3049 \\ 1.5183 \end{pmatrix}$

(1) $\Rightarrow y = 0.305 + 1.518x$ is the line of best fit.

2) $y = (1.5183) \times 8 + 0.305$

$= 12.45$ ice-creams

(so I know how much milk to buy tomorrow to make these ice-creams).

#.

HW Q) Repeat the above problem by assuming the model $y = a + bx + cx^2$ & compare the results.

** How would you pick a model? $y = a + bx$ vs $y = a + b\log(x) + c\sin x$?