Least Squares Regression

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

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

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

from Calculus.

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

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

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

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

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

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 $f(x) = x$, $g(x) = 0$.

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

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

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 $y = a + b\log(x) + c\sin x$?

Sketch of the derivation of the multi-dimensional least squares matrix-vector model:

PLAKSHA

the observables $\{y_i\}$ now depend "linearly" on more than one i/p features, say two i/p (features) dimensions - namely x_{2i} and x_{3i} (generally we do not use x_{1i} by convention as a symbol in this model)

data pts (say we have 10 of them)

$i=1$: $y_1 = \beta_1 + \beta_2 x_{21} + \beta_3 x_{31}$

$i=2$: $y_2 = \beta_1 + \beta_2 x_{22} + \beta_3 x_{32}$

$\vdots$

$i=10$: $y_{10} = \beta_1 + \beta_2 x_{210} + \beta_3 x_{310}$

In matrix vector form:

$$\vec{y} = \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{10} \end{pmatrix}}_{10\times 1} = \underbrace{\begin{pmatrix} 1 & x_{21} & x_{31} \\ 1 & x_{22} & x_{32} \\ \vdots & \vdots & \vdots \\ 1 & x_{210} & x_{310} \end{pmatrix}}_{10\times 3} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{3\times 1} = A\vec{\beta}$$

But A is not a square matrix;

$\therefore$ A is not invertible

$\therefore$ $\beta = A^{-1}\vec{y}$ is not a feasible calculation.

Essentially, this entails that a straight-line cannot be drawn through each of the ten data points. The line of "best fit" must be captured as $\vec{y} = A\bar{x} + \bar{e}$ where $\bar{e}$ must be as small as possible (minimized) to ensure a "best fit".

$$\bar{e} = (\bar{y} - A\bar{x})$$

$$E = \bar{e}^{+}\bar{e} = (\bar{y} - A\bar{x})^{+}(\bar{y} - A\bar{x}) = \sum_{i=1}^{10} e_i^2$$

Here $\bar{x}$ is the vector $\bar{\beta}$ (model parameters stacked as a vector)

$$\frac{\partial E}{\partial \bar{\beta}} = 0 = \frac{\partial}{\partial \bar{\beta}}\left\{\bar{y}^{+}\bar{y} - \bar{y}^{+}A\bar{\beta} - \bar{\beta}^{+}A^{+}\bar{y} + \bar{\beta}^{+}A^{+}A\bar{\beta}\right\}$$

Note $(AB)^{+} = B^{+}A^{+}$

Do you know the meaning of this term?

$$\Rightarrow -2A^{+}\bar{y} + 2A^{+}A\bar{\beta} = 0$$

Solving for $\bar{\beta}$,

$$\bar{\beta} = (A^{+}A)^{-1}A^{+}\bar{y}$$

is the least-squares soln. of the regression model.

Question: (I) Can you tell me a condition when $(A^{+}A)$ is not invertible?

(II) What do we do when $(A^{+}A)^{-1}$ does not exist?

DIFFERENTIATION WITH RESPECT TO A VECTOR

The first derivative of a scalar-valued function $f(\mathbf{x})$ with respect to a vector $\mathbf{x} = [x_1 \ x_2]^T$ is called the gradient of $f(\mathbf{x})$ and defined as

$$\nabla f(\mathbf{x}) = \frac{d}{d\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \partial f/\partial x_1 \\ \partial f/\partial x_2 \end{bmatrix} \tag{C.1}$$

Based on this definition, we can write the following equation.

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T \mathbf{y} = \frac{\partial}{\partial \mathbf{x}} \mathbf{y}^T \mathbf{x} = \frac{\partial}{\partial \mathbf{x}} (x_1 y_1 + x_2 y_2) = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{y} \tag{C.2}$$

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T \mathbf{x} = \frac{\partial}{\partial \mathbf{x}} (x_1^2 + x_2^2) = 2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2\mathbf{x} \tag{C.3}$$

Also with an $M \times N$ matrix A, we have

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T A \mathbf{y} = \frac{\partial}{\partial \mathbf{x}} \mathbf{y}^T A^T \mathbf{x} = A\mathbf{y} \tag{C.4a}$$

$$\frac{\partial}{\partial \mathbf{x}} \mathbf{y}^T A \mathbf{x} = \frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T A^T \mathbf{y} = A^T \mathbf{y} \tag{C.4b}$$

where

$$\mathbf{x}^T A \mathbf{y} = \sum_{m=1}^{M} \sum_{n=1}^{N} a_{mn} x_m y_n \tag{C.5}$$

Especially for a square, symmetric matrix A with $M = N$, we have

$$\frac{\partial}{\partial \mathbf{x}}\mathbf{x}^T A\mathbf{x} = (A + A^T)\mathbf{x} \xrightarrow{\text{if } A \text{ is symmetric}} 2A\mathbf{x} \tag{C.6}$$

The second derivative of a scalar function $f(\mathbf{x})$ with respect to a vector $\mathbf{x} = [x_1\ x_2]^T$ is called the Hessian of $f(\mathbf{x})$ and is defined as

$$H(\mathbf{x}) = \nabla^2 f(\mathbf{x}) = \frac{d^2}{d\mathbf{x}^2} f(\mathbf{x}) = \begin{bmatrix} \partial^2 f/\partial x_1^2 & \partial^2 f/\partial x_1 \partial x_2 \\ \partial^2 f/\partial x_2 \partial x_1 & \partial^2 f/\partial x_2^2 \end{bmatrix} \tag{C.7}$$

Based on this definition, we can write the following equation:

$$\frac{d^2}{d\mathbf{x}^2}\mathbf{x}^T A\mathbf{x} = A + A^T \xrightarrow{\text{if } A \text{ is symmetric}} 2A \tag{C.8}$$

On the other hand, the first derivative of a vector-valued function $\mathbf{f}(\mathbf{x})$ with respect to a vector $\mathbf{x} = [x_1\ x_2]^T$ is called the Jacobian of $f(\mathbf{x})$ and is defined as

$$J(\mathbf{x}) = \frac{d}{d\mathbf{x}}\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 \end{bmatrix} \tag{C.9}$$