X data matrix

N rows (samples)

$K+1$ columns (feature)

$K+1^{st}$ column is $\begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$

Y target data $N \times 1$ column vector

Goal: find $K+1 \times 1$ vector $M = \begin{pmatrix} M_1 \\ \vdots \\ M_{K+1} \end{pmatrix}$

so that

$$E = \|Y - XM\|^2 \text{ is minimized.}$$

$$MSE(M_1, \ldots, M_{K+1}) = \frac{1}{N} E = \frac{1}{N} \sum_{i=1}^{N} \Big(y_i - \underbrace{\sum_{j=1}^{K+1} x_{ij} M_j}_{\|Y - XM\|^2} \Big)^2$$

$K+1$ variables $M_1, \ldots, M_{K+1}$

$$\nabla E = \begin{bmatrix} \partial E / \partial M_1 \\ \partial E / \partial M_2 \\ \vdots \\ \partial E / \partial M_{K+1} \end{bmatrix} = 0$$

$A = (a_{ij})$ $r \times s$

$B = (b_{ij})$ $s \times t$

$C = (C_{ij})$ $r \times t$

$$c_{ij} = \sum_{k=1}^{s} a_{ik} b_{kj}.$$

$A = (a_{ij})$

$A^T = (a_{ji})$ $s \times r$ matrix

$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ $A^T = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$

Proposition: The gradient of $MSE(M) = E$ is given by

$E = \|Y - XM\|^2$

$X \quad N \times (K+1) \qquad M \; (K+1)\times 1$

$Y \quad N \times 1$

$$\nabla E = \begin{bmatrix} \frac{\partial}{\partial M_1}E \\ \frac{\partial}{\partial M_2}E \\ \vdots \\ \frac{\partial}{\partial M_{K+1}}E \end{bmatrix} = -2X^\intercal Y + 2X^\intercal XM \tag{5}$$

$= -2\left[X^T Y - X^T X M\right]$

$(K+1)\times 1$ matrix

where $X^\intercal$ is the transpose of X.

Proof: First, remember that the ij entry of $X^\intercal$ is the ji entry of X. Also, we will use the notation $X[j,:]$ to mean the j^{th} row of X and $X[:,i]$ to mean the i^{th} column of X. (This is copied from the Python programming language; the ':' means that index runs over all possibilities).

$2(Y_j - \sum_{1}^{k+1} x_{js}M_s)[-X_{jt}]$

$X_{jt}M_t$

$1 \le t \le K+1$

Since

$$E = \sum_{j=1}^{N}(Y_j - \sum_{s=1}^{k+1} X_{js}M_s)^2$$

we compute:

$$\begin{aligned}
\frac{\partial}{\partial M_t}E &= -2\sum_{j=1}^{N} X_{jt}(Y_j - \sum_{s=1}^{k+1} X_{js}M_s) \\
&= -2(\sum_{j=1}^{N} Y_j X_{jt} - \sum_{j=1}^{N}\sum_{s=1}^{k+1} X_{jt}X_{js}M_s) \\
&= -2(\sum_{j=1}^{N} X^\intercal_{tj} Y_j - \sum_{j=1}^{N}\sum_{s=1}^{k+1} X^\intercal_{tj}X_{js}M_s) \\
&= -2(X^\intercal[t,:]Y - \sum_{s=1}^{k+1}\boxed{\sum_{j=1}^{N} X^\intercal_{tj}X_{js}}M_s) \\
&= -2(X^\intercal[t,:]Y - \boxed{\sum_{s=1}^{k+1}(X^\intercal X)_{ts}M_s}) \\
&= -2(X^\intercal[t,:]Y - (X^\intercal X)[t,:]M)
\end{aligned} \tag{6}$$

$\sum X^T_{tj} Y_j$

t^{th} row $\begin{pmatrix} x_{t1} & x_{t2} & \cdots \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}$

X^T

$X^T[t,:]$

t^{th} row of X^T

t^{th} entry of ∇E is this formula

Stacking up the different rows to make E yields the desired formula.

$$\nabla E = -2(X^T Y - X^T X M) = 0$$

$$\underbrace{X^T X}_{K+1 \times K+1} \underset{K+1 \times 1}{M} = X^T Y \qquad K+1 \times 1$$

Assume

D is invertible

$$D = X^T X \quad (K+1) \times (K+1)$$

$$M = D^{-1} X^T Y$$

$$Y_{predicted} = XM$$

$$E = \|Y - XM\|^2 = \|Y - Y_{predicted}\|^2$$

$$Y_{predicted} = \underbrace{\overset{N \times K+1}{X} \; \overset{K+1 \times K+1}{D^{-1}} \; \overset{K+1 \times N}{X^T} \; \overset{N \times 1}{Y}}_{N \times 1}$$

predicted

$M \Leftrightarrow m, b$

$y = mx + b$

$Y \quad N \times 1$