X data matrix N rows, 1 for each sample
K+1 columns
K features
Last column is $\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$

Y target values $N \times 1$ column vector

M $(K+1)\times 1$ column vector m_{K+1} is the intercept

Minimize $MSE = \frac{1}{N} \|Y - XM\|^2$

$Y - XM$ is an $N \times 1$ vector

$$Y - XM = \left(y_i - \sum_{j=1}^{K+1} x_{ij} m_j \right)$$

$$\|Y - XM\|^2 = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{K+1} x_{ij} m_j \right)^2$$

minimize this

m_j are variables

$A = (a_{ij})$ which is $r \times s$
$B = (b_{ij})$ which is $s \times t$

$C = AB$ is $r \times t$

$$c_{ik} = \sum_{j=1}^{s} a_{ij} b_{jk}$$

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

Minimize MSE by taking its $\frac{\partial}{\partial M_i}$ $i = 1, \ldots, k+1$ and setting $= 0$.

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

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

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

$$= -2\left[\underbrace{X^\intercal Y}_{(k+1)\times 1} - \underbrace{X^\intercal X M}_{(k+1)\times 1}\right]$$

X $N \times (k+1)$

$X^\intercal$ $(k+1) \times N$

Y $N \times 1$

$X^\intercal X = (k+1) \times (k+1)$

M $(k+1) \times 1$

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).

Since

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

we compute:

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

$X^\intercal[t,:]$ t^{th} row of $X^\intercal$

$(X^\intercal X)[t,:]$ t^{th} row of $X^\intercal X$

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

$$\nabla E = -2\left[X^\intercal Y - X^\intercal X M\right].$$

$$\nabla E = 0$$

$$\underbrace{X^\intercal Y}_{(k+1\times N)\, N\times 1} = \underbrace{X^\intercal X}_{(k+1)\times(k+1)} \underbrace{M}_{(k+1)\times 1}$$

$$D = X^\intercal X \quad (k+1)\times(k+1)$$

Suppose D is an invertible matrix.

$$\boxed{M = D^{-1} X^\intercal Y}$$