

Multiple Linear Regression - Normal Equation

Ugenteraan Manogaran

E-mail : m.ugenteraan15@gmail.com

April 16, 2019

Suppose the inputs are :

$$\mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & . & . & . & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & . & . & . & x_n^{(2)} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_1^{(m)} & x_2^{(m)} & . & . & . & x_n^{(m)} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ . \\ . \\ . \\ y^{(m)} \end{bmatrix} \quad (1)$$

where each row in \mathbf{X} is the i -th sample. Each column in \mathbf{X} represents the feature (dependent variable) of the dataset.

The goal is to find a linear function \mathbf{h} to approximate $y^{(i)}$, given $\mathbf{x}^{(i)}$

$x_0^{(i)}$ will be added into \mathbf{X} where $x_0^{(i)} = 1$ to simplify the notations for the finding of the constant in the linear equation later. Hence,

$$\mathbf{X} = \begin{bmatrix} x_0^{(1)} & x_1^{(1)} & . & . & . & x_n^{(1)} \\ x_0^{(2)} & x_1^{(2)} & . & . & . & x_n^{(2)} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ x_0^{(m)} & x_1^{(m)} & . & . & . & x_n^{(m)} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ . \\ . \\ . \\ y^{(m)} \end{bmatrix} \quad (2)$$

The linear function \mathbf{h} is

$$\mathbf{h}_\theta(\mathbf{x}^{(i)}) = \theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \dots + \theta_n x_n^{(i)} \quad (3)$$

or

$$\mathbf{h}_\theta(\mathbf{x}^{(i)}) = \sum_{j=0}^n \theta_j x_j^{(i)} \quad (4)$$

where $\theta_i \in \mathbb{R}$ and $i = 1, \dots, m$, such that

$$\mathbf{J}(\theta_0, \dots, \theta_n) = \frac{1}{2m} \sum_{i=1}^m (\mathbf{h}_\theta(\mathbf{x}^{(i)}) - y^{(i)})^2 \quad (5)$$

is minimized.

Taking θ as a vector,

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \cdot \\ \cdot \\ \cdot \\ \theta_n \end{bmatrix} \quad (6)$$

we can rewrite (3) or (4) as

$$\mathbf{h}_\theta(\mathbf{x}^{(i)}) = \boldsymbol{\theta}^T \mathbf{x}^{(i)} \quad (7)$$

Since

$$(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T (\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) = \sum_{i=1}^m (\mathbf{h}_\theta(\mathbf{x}^{(i)}) - y^{(i)})^2 \quad , \quad (8)$$

(5) can be rewritten as (Note that $\mathbf{X}\boldsymbol{\theta}$ is a vector.)

$$\mathbf{J}(\boldsymbol{\theta}) = \frac{1}{2m} (\boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\theta} - 2(\mathbf{X}\boldsymbol{\theta})^T \mathbf{y} + \mathbf{y}^T \mathbf{y}) \quad \text{See Appendix 1.1} \quad (9)$$

Since \mathbf{J} is a polynomial function of degree 2, to find $\boldsymbol{\theta}$ such that \mathbf{J} is at minimum, we want to find

$$\frac{\partial \mathbf{J}}{\partial \boldsymbol{\theta}} = 0 \quad (10)$$

Since,

$$\frac{\partial \mathbf{J}}{\partial \boldsymbol{\theta}} = 2\mathbf{X}^T \mathbf{X} \boldsymbol{\theta} - 2\mathbf{X}^T \mathbf{y} \quad \text{See Appendix 1.2} \quad (11)$$

then,

$$\begin{aligned} 2\mathbf{X}^T \mathbf{X} \boldsymbol{\theta} - 2\mathbf{X}^T \mathbf{y} &= 0 \\ \mathbf{X}^T \mathbf{X} \boldsymbol{\theta} &= \mathbf{X}^T \mathbf{y} \end{aligned} \quad (12)$$

This system (12) is known as the **normal equations** for $\boldsymbol{\theta}$. Furthermore, if $\mathbf{X}^T \mathbf{X}$ is invertible, then there exist a unique solution for $\boldsymbol{\theta}$, such that

$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (13)$$

References

- [1] *Nakos, G., and Joyner, D. (1998). Linear algebra with applications. PWS Publishing Company.*
- [2] *Derivation of the Normal Equation for linear regression - Eli Bendersky's website, 2019*

1 Appendix

1.1 Proof for (9)

Using equation (5) and (8),

$$\mathbf{J}(\boldsymbol{\theta}) = \frac{1}{2m}(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})$$

Note : $(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})$ is a vector. To multiply a vector by its own tranpose is equivalent to squaring the vector.

$$\begin{aligned} &= \frac{1}{2m}((\mathbf{X}\boldsymbol{\theta})^T - \mathbf{y}^T)(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) \\ &= \frac{1}{2m}(\boldsymbol{\theta}^T \mathbf{X}^T - \mathbf{y}^T)(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}) \\ &= \frac{1}{2m}(\boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X}\boldsymbol{\theta} - (\mathbf{X}\boldsymbol{\theta})^T \mathbf{y} - \mathbf{y}^T (\mathbf{X}\boldsymbol{\theta}) + \mathbf{y}^T \mathbf{y}) \end{aligned}$$

Since $(\mathbf{X}\boldsymbol{\theta})^T \mathbf{y} = \mathbf{y}^T (\mathbf{X}\boldsymbol{\theta})$, See Appendix 1.1.1

$$\mathbf{J}(\boldsymbol{\theta}) = \frac{1}{2m}(\boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X}\boldsymbol{\theta} - 2(\mathbf{X}\boldsymbol{\theta})^T \mathbf{y} + \mathbf{y}^T \mathbf{y})$$

1.1.1 Proof for $(\mathbf{X}\boldsymbol{\theta})^T \mathbf{y} = \mathbf{y}^T (\mathbf{X}\boldsymbol{\theta})$

$$\begin{aligned}
(\mathbf{X}\boldsymbol{\theta})^T \mathbf{y} &= \left(\begin{bmatrix} x_0^{(1)} & \cdot & \cdot & \cdot & x_n^{(1)} \\ \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ x_0^{(m)} & \cdot & \cdot & \cdot & x_n^{(m)} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \cdot \\ \cdot \\ \cdot \\ \theta_n \end{bmatrix} \right)^T \begin{bmatrix} y^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ y^{(m)} \end{bmatrix} \\
&= \left(\begin{bmatrix} x_0^{(1)}\theta_0 & + & \cdot & \cdot & \cdot & + & x_n^{(1)}\theta_n \\ \cdot & & \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot & & \cdot \\ x_0^{(m)}\theta_0 & + & \cdot & \cdot & \cdot & + & x_n^{(m)}\theta_n \end{bmatrix} \right)^T \begin{bmatrix} y^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ y^{(m)} \end{bmatrix} \\
&= \begin{bmatrix} (x_0^{(1)}\theta_0)y^{(1)} & + & \cdot & \cdot & \cdot & + & (x_0^{(m)}\theta_0)y^{(m)} \\ \cdot & & \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot & & \cdot \\ (x_n^{(1)}\theta_n)y^{(1)} & + & \cdot & \cdot & \cdot & + & (x_n^{(m)}\theta_n)y^{(m)} \\ y^{(1)}(x_0^{(1)}\theta_0) & + & \cdot & \cdot & \cdot & + & y^{(m)}(x_0^{(m)}\theta_0) \\ \cdot & & \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot & & \cdot \\ y^{(1)}(x_n^{(1)}\theta_n) & + & \cdot & \cdot & \cdot & + & y^{(m)}(x_n^{(m)}\theta_n) \end{bmatrix} \\
&= \left(\begin{bmatrix} y^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ y^{(m)} \end{bmatrix} \right)^T \left(\begin{bmatrix} x_0^{(1)}\theta_0 & + & \cdot & \cdot & \cdot & + & x_n^{(1)}\theta_n \\ \cdot & & \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot & & \cdot \\ x_0^{(m)}\theta_0 & + & \cdot & \cdot & \cdot & + & x_n^{(m)}\theta_n \end{bmatrix} \right) \\
&= \begin{bmatrix} y^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ y^{(m)} \end{bmatrix}^T \left(\begin{bmatrix} x_0^{(1)} & \cdot & \cdot & \cdot & x_n^{(1)} \\ \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ x_0^{(m)} & \cdot & \cdot & \cdot & x_n^{(m)} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \cdot \\ \cdot \\ \cdot \\ \theta_n \end{bmatrix} \right) \\
&= \mathbf{y}^T (\mathbf{X}\boldsymbol{\theta})
\end{aligned}$$

1.2 Proof for (11)

$$\mathbf{J}(\boldsymbol{\theta}) = \frac{1}{2m} (\boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\theta} - 2(\mathbf{X}\boldsymbol{\theta})^T \mathbf{y} + \mathbf{y}^T \mathbf{y})$$

The equation above (9) be broken down to three expressions,

$$\boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\theta} \tag{14}$$

,

$$2(\mathbf{X}\boldsymbol{\theta})^T \mathbf{y} \tag{15}$$

and

$$\mathbf{y}^T \mathbf{y} \quad (16)$$

Expression (16) does not depend on $\boldsymbol{\theta}$, hence the partial derivative of the expression with respect to $\boldsymbol{\theta}$ results in 0.

As for expression (14),

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} (\boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \left(\begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix}^T \begin{bmatrix} x_0^{(1)} & \cdot & \cdot & \cdot & x_n^{(1)} \\ \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ x_0^{(m)} & \cdot & \cdot & \cdot & x_n^{(m)} \end{bmatrix}^T \begin{bmatrix} x_0^{(1)} & \cdot & \cdot & \cdot & x_n^{(1)} \\ \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ x_0^{(m)} & \cdot & \cdot & \cdot & x_n^{(m)} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix} \right) \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \left(\begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix}^T \begin{bmatrix} (x_0^{(1)} x_0^{(1)} + \dots + x_0^{(m)} x_0^{(m)}) & \cdot & \cdot & \cdot & (x_0^{(1)} x_n^{(1)} + \dots + x_0^{(m)} x_n^{(m)}) \\ \cdot & \cdot & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ (x_n^{(1)} x_0^{(1)} + \dots + x_n^{(m)} x_0^{(m)}) & \cdot & \cdot & \cdot & (x_n^{(1)} x_n^{(1)} + \dots + x_n^{(m)} x_n^{(m)}) \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix} \right) \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \left(\begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix}^T \begin{bmatrix} (x_0^{(1)} x_0^{(1)} + \dots + x_0^{(m)} x_0^{(m)}) \theta_0 + \dots + (x_0^{(1)} x_n^{(1)} + \dots + x_0^{(m)} x_n^{(m)}) \theta_n \\ \cdot \\ \cdot \\ \cdot \\ (x_n^{(1)} x_0^{(1)} + \dots + x_n^{(m)} x_0^{(m)}) \theta_0 + \dots + (x_n^{(1)} x_n^{(1)} + \dots + x_n^{(m)} x_n^{(m)}) \theta_n \end{bmatrix} \right) \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} (\theta_0 ((x_0^{(1)} x_0^{(1)} + \dots + x_0^{(m)} x_0^{(m)}) \theta_0 + \dots + (x_0^{(1)} x_n^{(1)} + \dots + x_0^{(m)} x_n^{(m)}) \theta_n) + \dots + \\ &\quad \theta_n ((x_n^{(1)} x_0^{(1)} + \dots + x_n^{(m)} x_0^{(m)}) \theta_0 + \dots + (x_n^{(1)} x_n^{(1)} + \dots + x_n^{(m)} x_n^{(m)}) \theta_n)) \end{aligned}$$

Since,

$$\begin{aligned} \frac{\partial}{\partial \theta_0} (\boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\theta}) &= [2(x_0^{(1)} x_0^{(1)} + \dots + x_0^{(m)} x_0^{(m)}) \theta_0 + \dots + (x_n^{(1)} x_0^{(1)} + \dots + x_n^{(m)} x_0^{(m)}) \theta_n] \\ &\quad + \dots + [(x_0^{(1)} x_n^{(1)} + \dots + x_0^{(m)} x_n^{(m)}) \theta_n] \\ &= [2(x_0^{(1)} x_0^{(1)} + \dots + x_0^{(m)} x_0^{(m)}) \theta_0 + \dots + 2(x_n^{(1)} x_0^{(1)} + \dots + x_n^{(m)} x_0^{(m)}) \theta_n] \\ &= 2[(x_0^{(1)} x_0^{(1)} + \dots + x_0^{(m)} x_0^{(m)}) \theta_0 + \dots + (x_n^{(1)} x_0^{(1)} + \dots + x_n^{(m)} x_0^{(m)}) \theta_n] \\ &\quad \cdot \\ &\quad \cdot \\ &\quad \cdot \\ \frac{\partial}{\partial \theta_n} (\boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\theta}) &= [(x_n^{(1)} x_0^{(1)} + \dots + x_n^{(m)} x_0^{(m)}) \theta_0] + \dots + [(x_0^{(1)} x_n^{(1)} + \dots + x_0^{(m)} x_n^{(m)}) \theta_0 \\ &\quad + \dots + 2(x_n^{(1)} x_n^{(1)} + \dots + x_n^{(m)} x_n^{(m)}) \theta_n] \\ &= [2(x_n^{(1)} x_0^{(1)} + \dots + x_n^{(m)} x_0^{(m)}) \theta_0] + \dots + 2(x_n^{(1)} x_n^{(1)} + \dots + x_n^{(m)} x_n^{(m)}) \theta_n] \\ &= 2[(x_n^{(1)} x_0^{(1)} + \dots + x_n^{(m)} x_0^{(m)}) \theta_0] + \dots + (x_n^{(1)} x_n^{(1)} + \dots + x_n^{(m)} x_n^{(m)}) \theta_n] \end{aligned}$$

then,

$$\frac{\partial}{\partial \boldsymbol{\theta}} (\boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\theta}) = 2 \mathbf{X}^T \mathbf{X} \boldsymbol{\theta}$$

As for expression (15),

$$\frac{\partial}{\partial \boldsymbol{\theta}} (2(\mathbf{X}\boldsymbol{\theta})^T \mathbf{y}) = \frac{\partial}{\partial \boldsymbol{\theta}} \left(2 \begin{bmatrix} (x_0^{(1)} \theta_0) y^{(0)} & + & . & . & . & + & (x_0^{(m)} \theta_0) y^{(n)} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ (x_n^{(1)} \theta_n) y^{(0)} & + & . & . & . & + & (x_n^{(m)} \theta_n) y^{(n)} \end{bmatrix} \right)$$

Since

$$\begin{aligned} \frac{\partial}{\partial \theta_0} (2(\mathbf{X}\boldsymbol{\theta})^T \mathbf{y}) &= 2((x_0^{(1)}) y^{(0)} + \dots + (x_0^{(m)}) y^{(n)}) \\ &\quad . \\ &\quad . \\ &\quad . \\ \frac{\partial}{\partial \theta_n} (2(\mathbf{X}\boldsymbol{\theta})^T \mathbf{y}) &= 2((x_n^{(1)}) y^{(0)} + \dots + (x_n^{(m)}) y^{(n)}) \end{aligned}$$

then,

$$\frac{\partial}{\partial \boldsymbol{\theta}} (2(\mathbf{X}\boldsymbol{\theta})^T \mathbf{y}) = 2 \mathbf{X}^T \mathbf{y}$$