In the normal equation approach for linear regression, the normal equation is given by -

$$\beta = (X^T X)^{-1} \cdot (X^T y)$$

where $\beta \rightarrow$ hypothesis parameters

$X \rightarrow$ Input feature for each instance

$Y \rightarrow$ Output value of each instance

Derivation:

The hypothesis function is given by -

$$h(\theta) = \theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_n x_n$$

$$\Rightarrow h(\beta) = \beta^T X \rightarrow \text{Hypothesis function}$$

So, predicted value for the output variable $= \beta^T X$

Cost function, $J(\beta) = \sum (y_{predicted_i} - y_i)^2$

$$\Rightarrow \sum_{i=1}^{n} (\beta^T x_i - y_i)^2$$

$$\Rightarrow \sum_{i=1}^{n} (\beta^T x_i - y_i)^T (\beta^T x_i - y_i)$$

$$= (X\beta - y)^T (X\beta - y)$$

$n \rightarrow$ the no of data points/ samples.

$x_i \rightarrow$ the i^{th} data point

$y_i \rightarrow$ actual label for the i^{th} data point

To minimise the function

$$J(\beta) = (X\beta - y)^T (X\beta - y)$$

$$\frac{\partial J(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta} \left((X\beta - y)^T (X\beta - y) \right)$$

$$= 2X^T X\beta - 2X^T y = 0 \Rightarrow 2(X^T X\beta - X^T y) = 0$$

$$\Rightarrow \quad X^T X \beta = X^T y$$

$$(X^T X)^{-1}(X^T X)\beta = (X^T X)^{-1} X^T y$$

$$I\beta = (X^T X)^{-1} X^T y$$

$$\Rightarrow \boxed{\beta = (X^T X)^{-1} X^T y}$$

$\therefore$ Hence proved

Limitations:

1. The normal ~~the~~ equation method involves multiple matrix multiplication & a matrix inversion as well, which is very costly & takes $O(n^3)$ time to compute. It is very expensive computation for large ~~the~~ datasets.

2. ~~Second~~ It is very sensitive to outliers in the data as it minimizes the sum of squares b/w the actual & predicted value. So, before performing normal equation approach, we need to ~~p~~ remove outliers.