Suppose we have 3 data points on a graph where x is our feature and y is. our label.

$(x^1,\ y^1)(x^2,\ y^2)(x^3,\ y^3)$

For example x can be age can be a feature and our label y could be salary(USD). Now suppose we want to know the possible salary for x^* and we dont know the label but we want a possible salary for someone who is that age.

What we do with linear regression is seek a line that bests fits this data. what we want is a line that somehow approximates those points. So How can we find this line?

We want

What we want is to solve a system of equations. So for example, we want $y^1 = w_i x^1 + w_2$. We also want to hit the other point with the line, $y^2 = w_i x^2 + w_2$, and we also want to hit $y^3 = w_i x^3 + w_2$ with the line. But there is no line that will fit all those pooints. So in regression the best we can possibly do is approx those points.

$$y^1 \approxeq w_i x^1 + w_2$$

$$y^2 \approxeq w_i x^2 + w_2$$

$$y^3 \approxeq w_i x^3 + w_2$$

But let's ignore the approximation part of this for now. What does this system look like? Well it looks like this:

$$\begin{bmatrix} x^1 1 \\ x^2 1 \\ x^3 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} y^1 \\ y^2 \\ y^3 \end{bmatrix}$$

To solve this problem, we need to solve this system of equations. So we can denote this matrix as

$$\begin{bmatrix} x^1 1 \\ x^2 1 \\ x^3 1 \end{bmatrix} \Rightarrow X$$

and we can denote this vector as

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \Rightarrow w$$

and then this vector by

$$\begin{bmatrix} y^1 \\ y^2 \\ y^3 \end{bmatrix} \Rightarrow y$$

so what we're wishing to solve is $Xw = y$

X is likely non-invertible so the best we can do is solve for an approximation.Say $\hat{w}$ with $X\hat{w} \approx y$. That's the best we can do. So how do we find this?

Well starting from this this equation, we want $Xw = y$ and we know that X is more than likely not invertible. But we can play with this, and multiply both sides by X^T

$$X^T X\hat{w} = X^T y$$

so lets call this solution to this system $\hat{w}$

Now X^TX is going to be invertible. So what we have now is that,

$$X^TX\hat{w} = X^Ty \Rightarrow \hat{w} = (X^TX)^{-1}X^Ty$$

this here is the **closed-form solution** to linear regression.

Note: $(X^TX)^{-1}$ *is called the pseudo inverse of X*

So now thaat we solve this system of equations, we're going to say that if we want to know what the approximate label for this x^* is , then we'll say that $y \approx \hat{w}_1x^* + \hat{w}_2$ once we solve for $\hat{w}$. This is called a prediction. This prediction for a given input will give us an approximate output provided the data closely lies to some linear approximation.