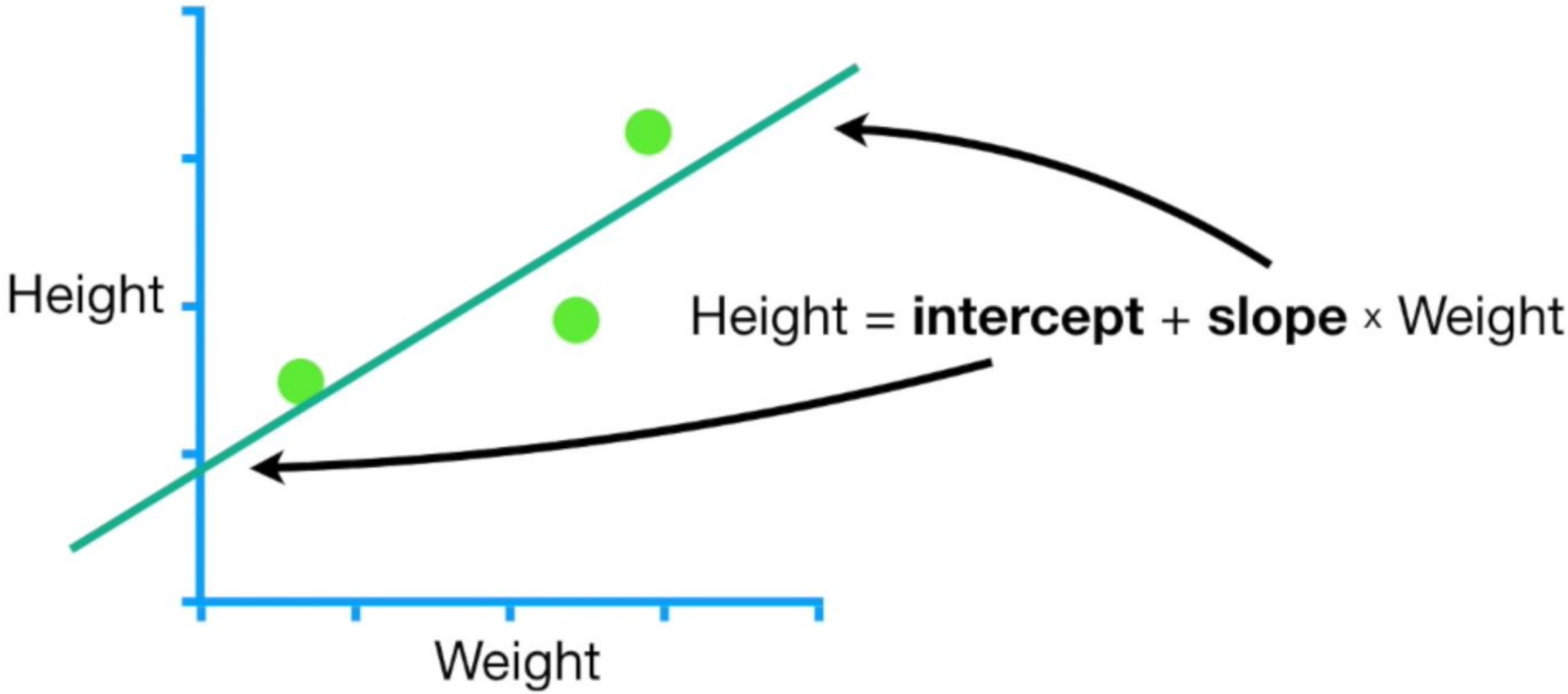
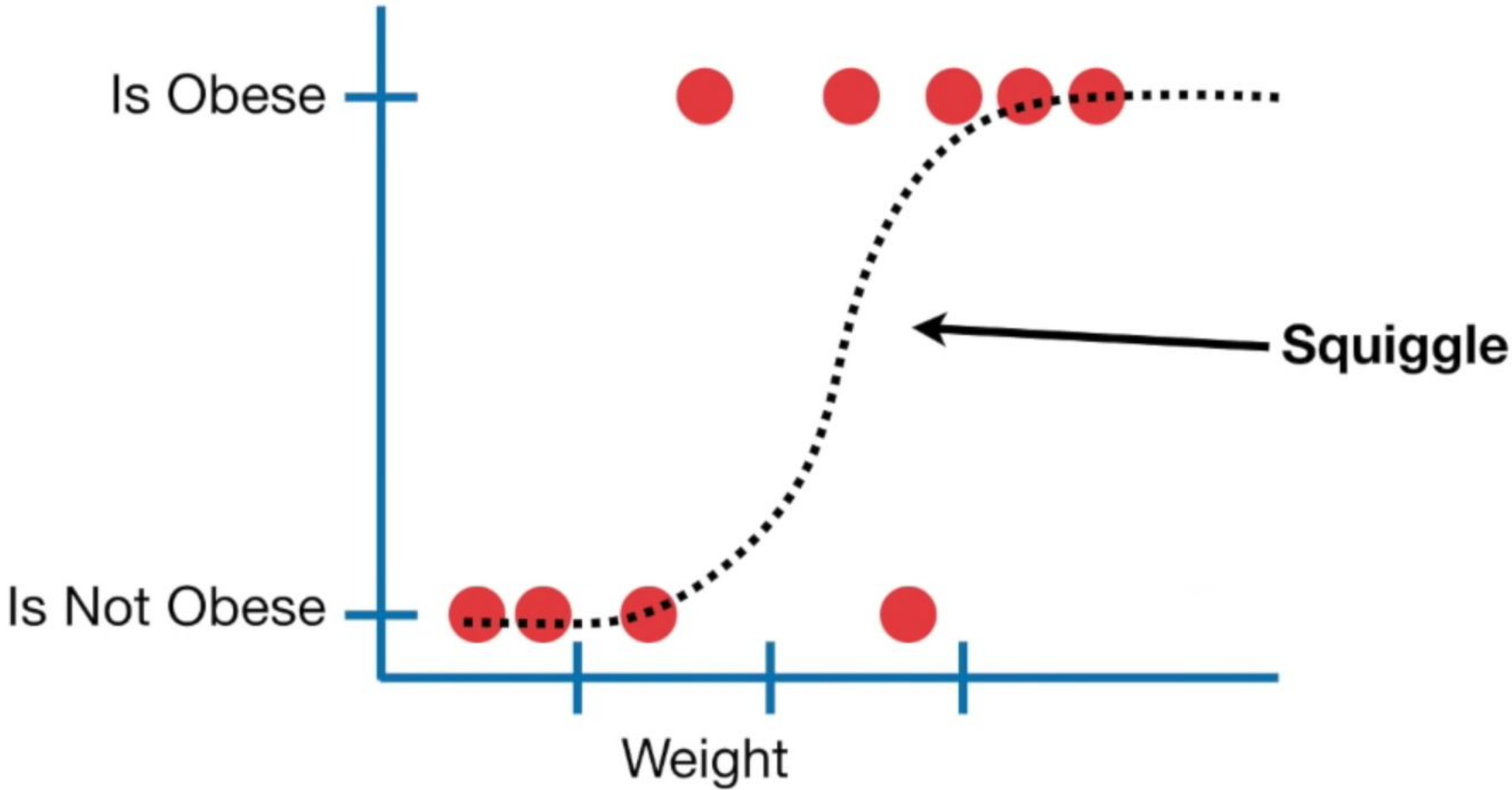
Prediction

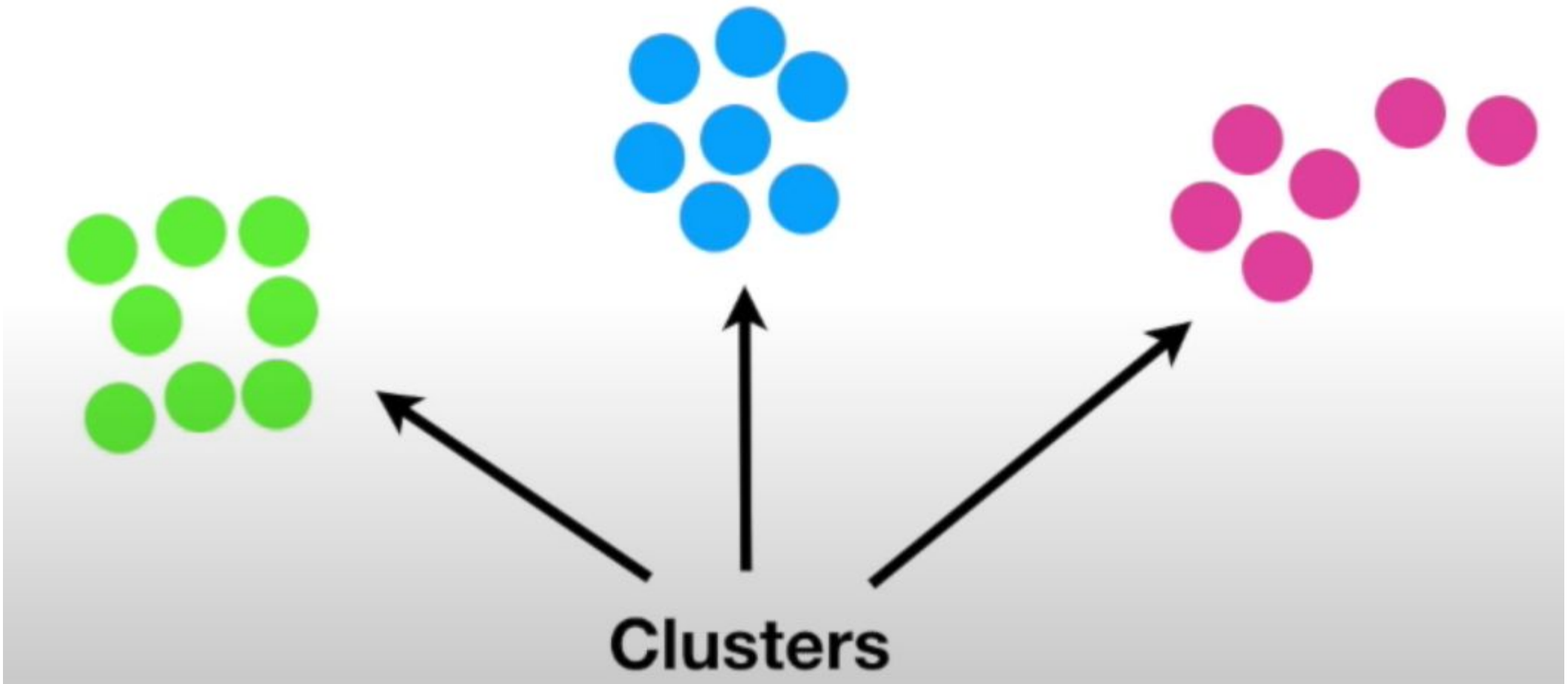
Error function, Convergence and Gradient Descent

When we fit a line with **Linear Regression**, we optimize the **Intercept** and **Slope**.

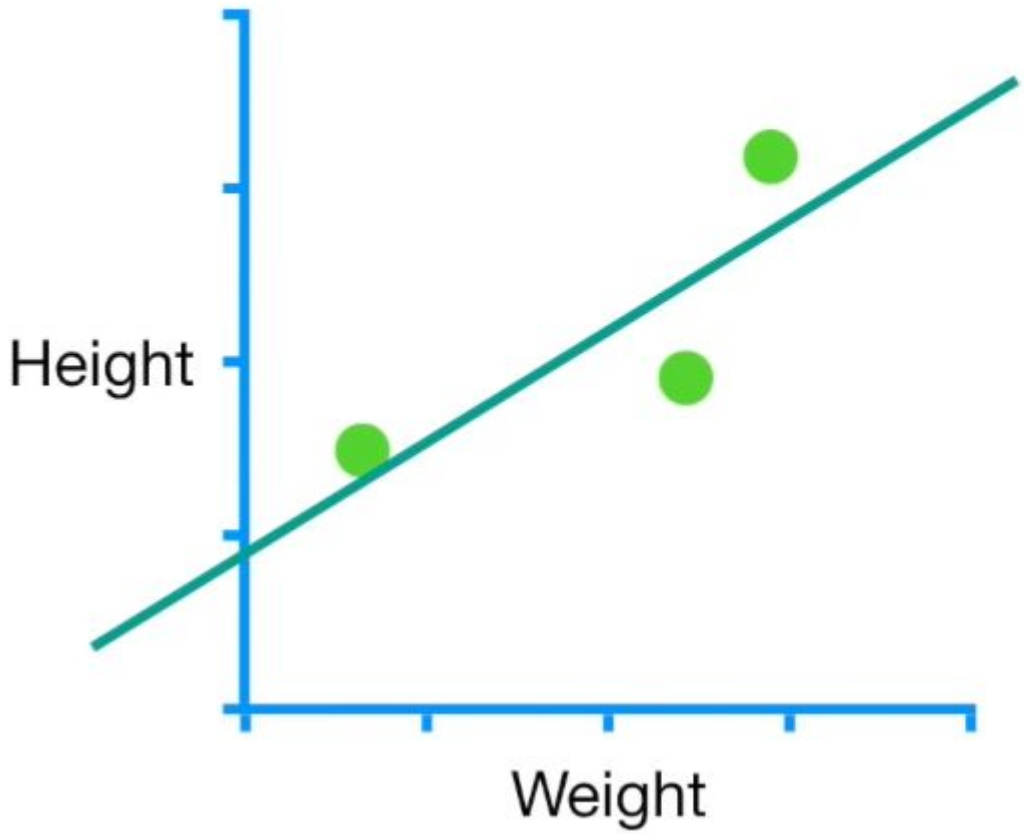


When we use **Logistic Regression**,
we optimize a squiggle.



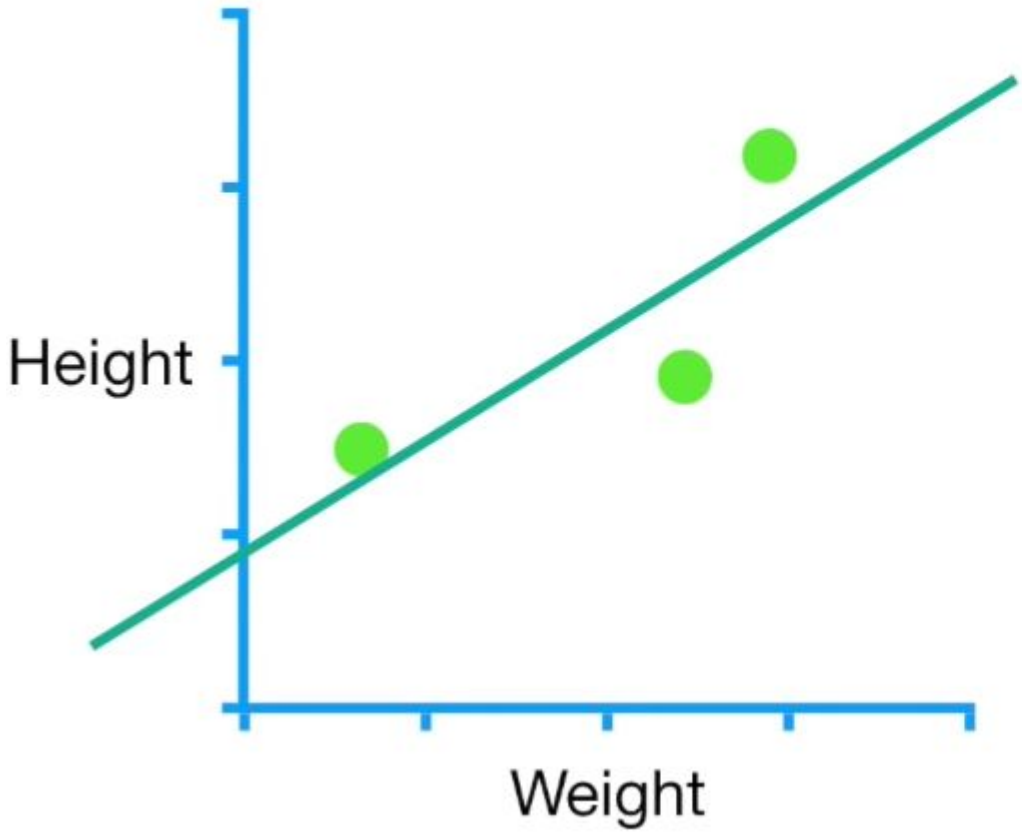

Clusters

Predicted Height = intercept + slope × **Weight**



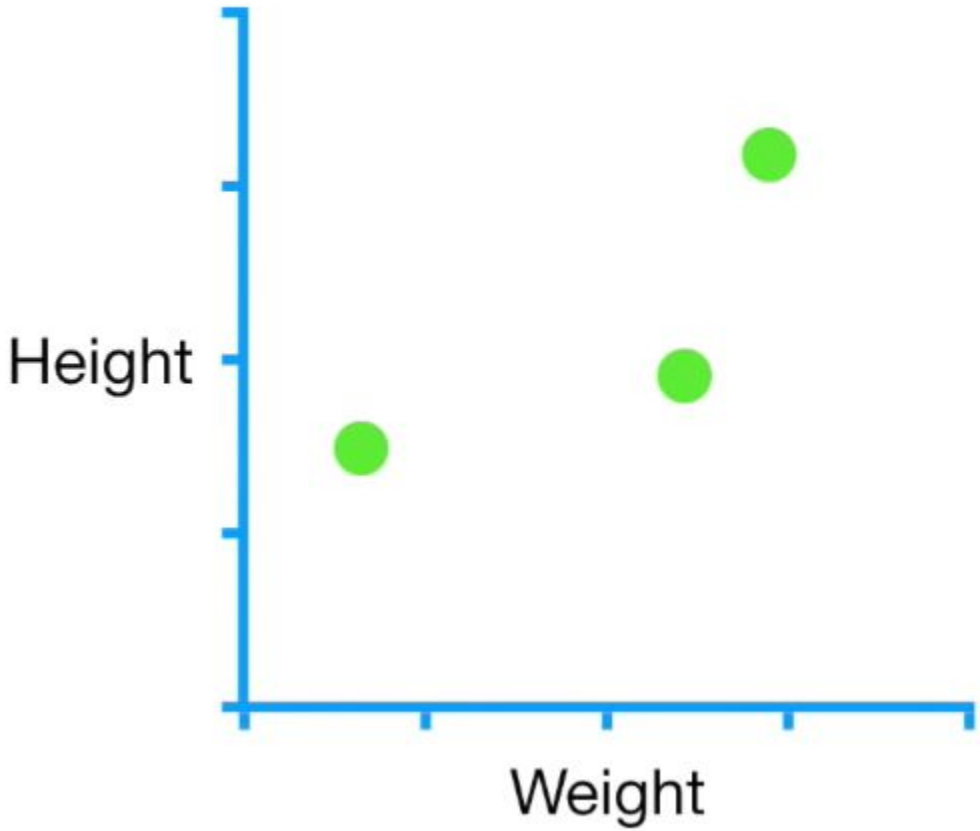
So let's learn how **Gradient Descent** can fit a line to data by finding the optimal values for the **Intercept** and the **Slope**.

Predicted Height = intercept + slope × **Weight**



So for now, let's just plug in the **Least Squares** estimate for the **Slope**, **0.64**.

Predicted Height = intercept + 0.64 × **Weight**

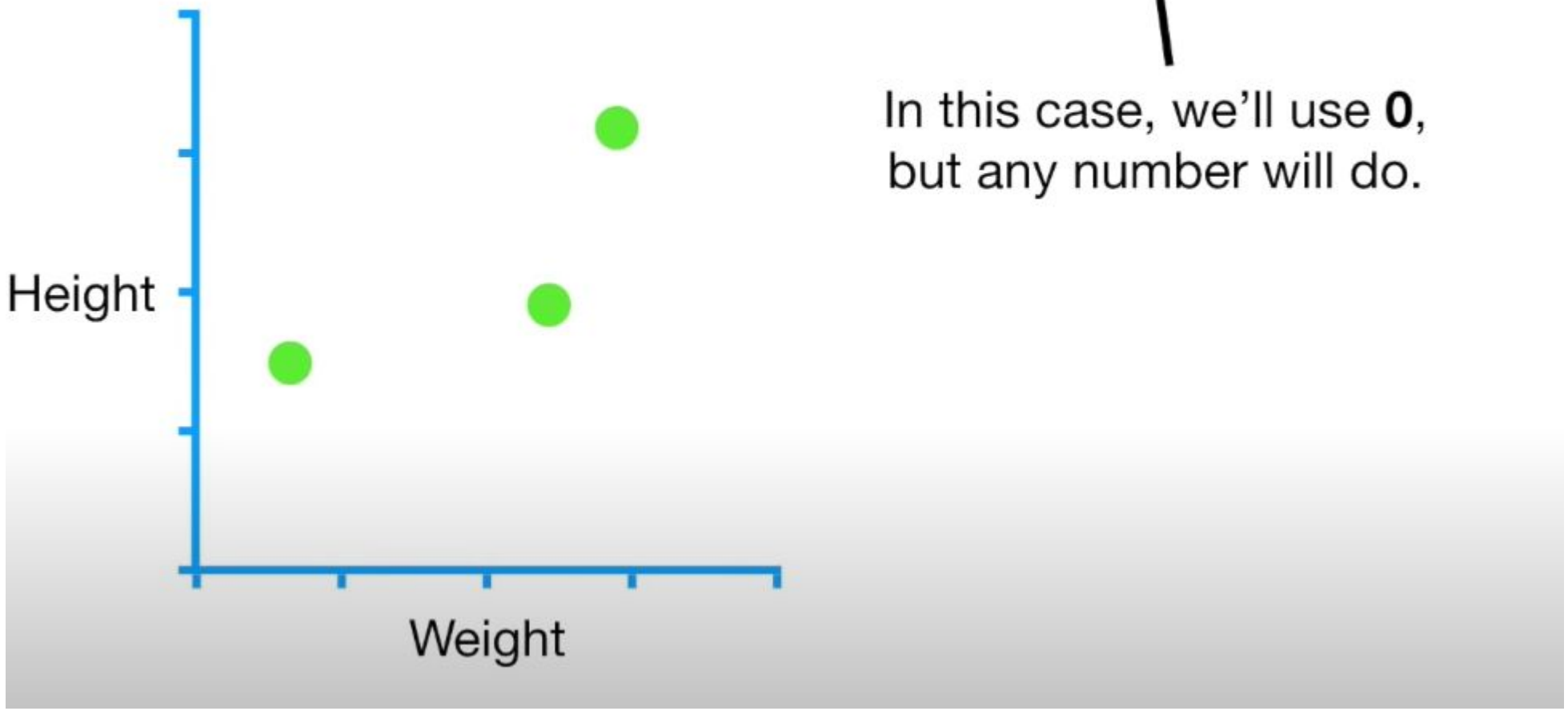


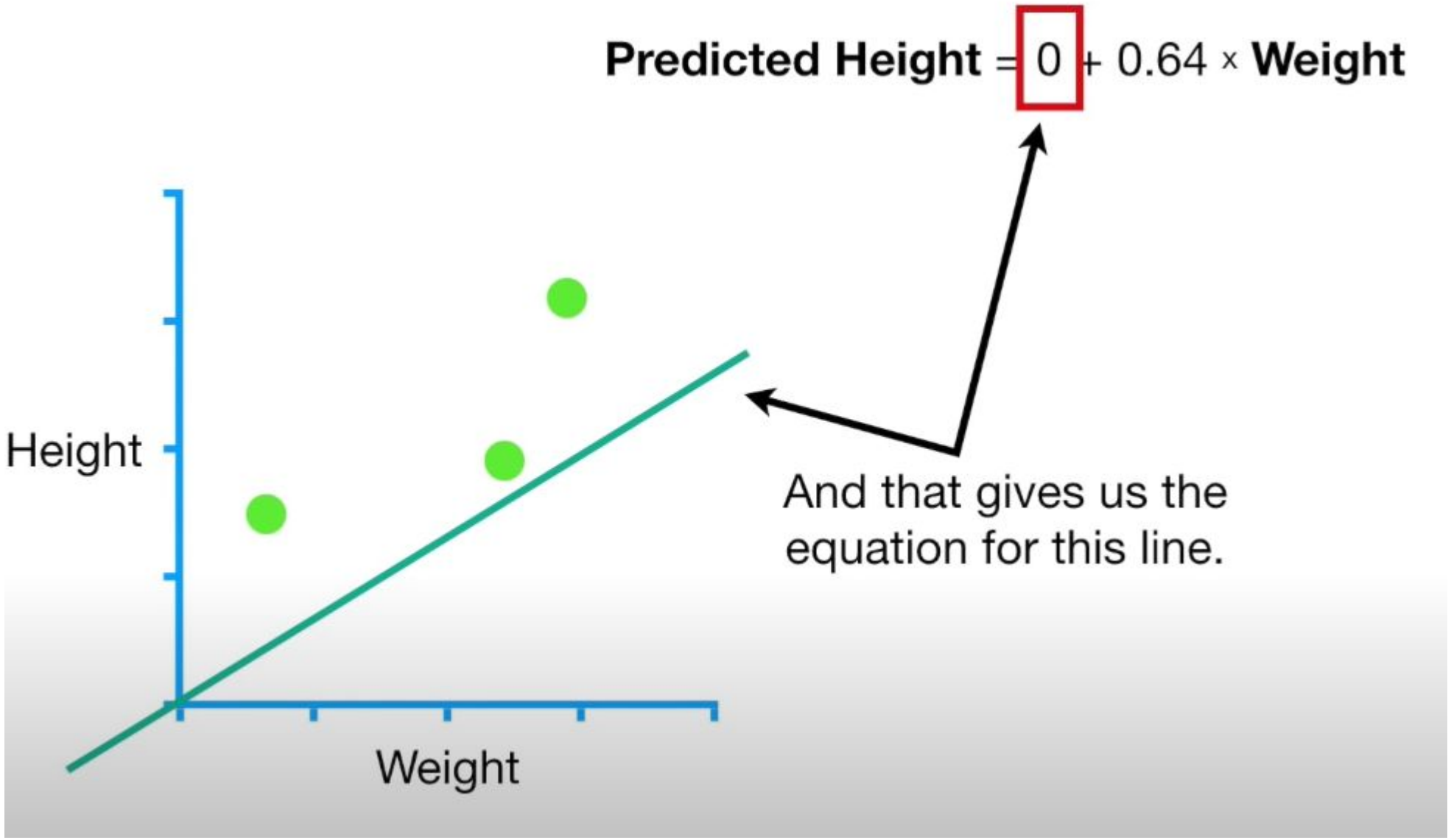
The first thing we do is pick a random value for the **Intercept**.

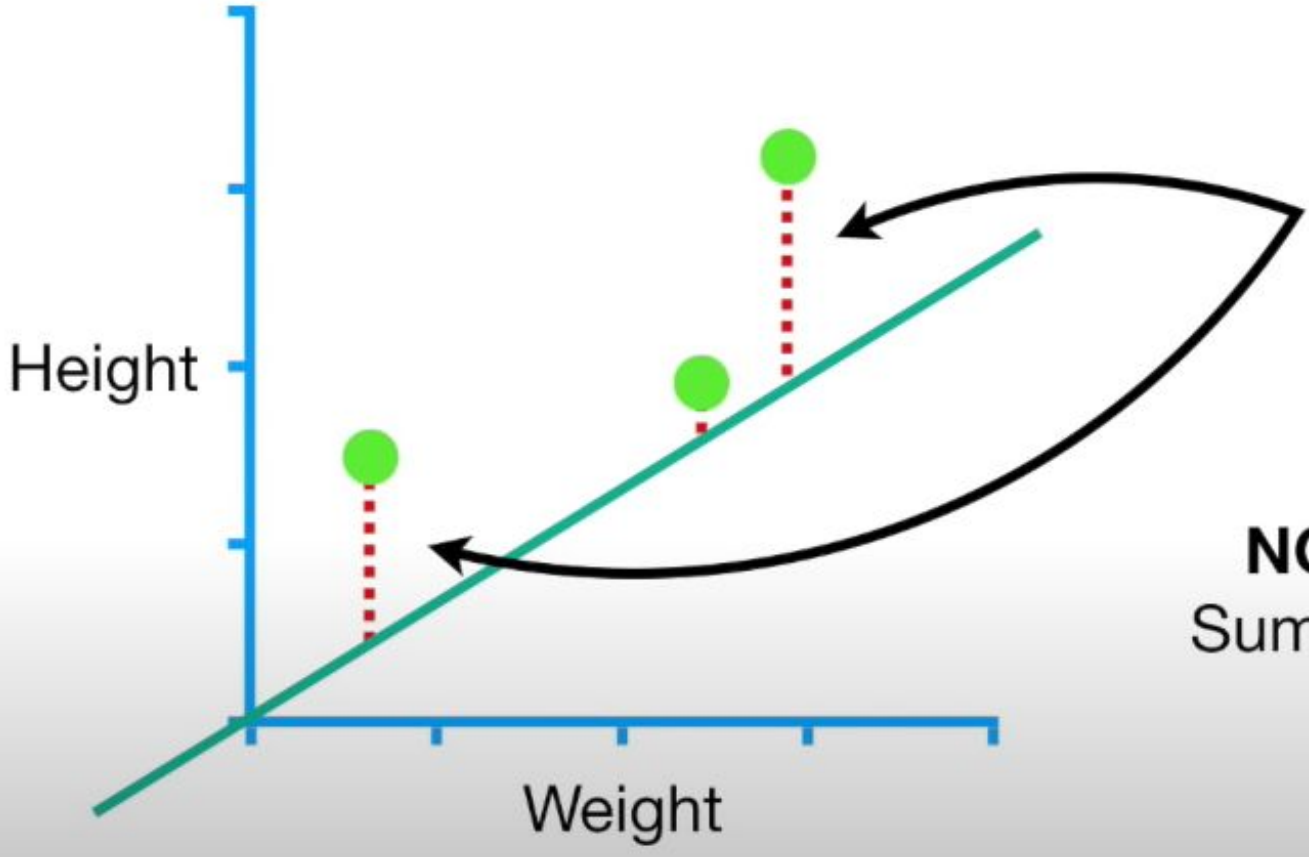
This is just an initial guess that gives **Gradient Descen**t something to improve upon.

Predicted Height = 0 + 0.64 × **Weight**

In this case, we'll use **0**,
but any number will do.

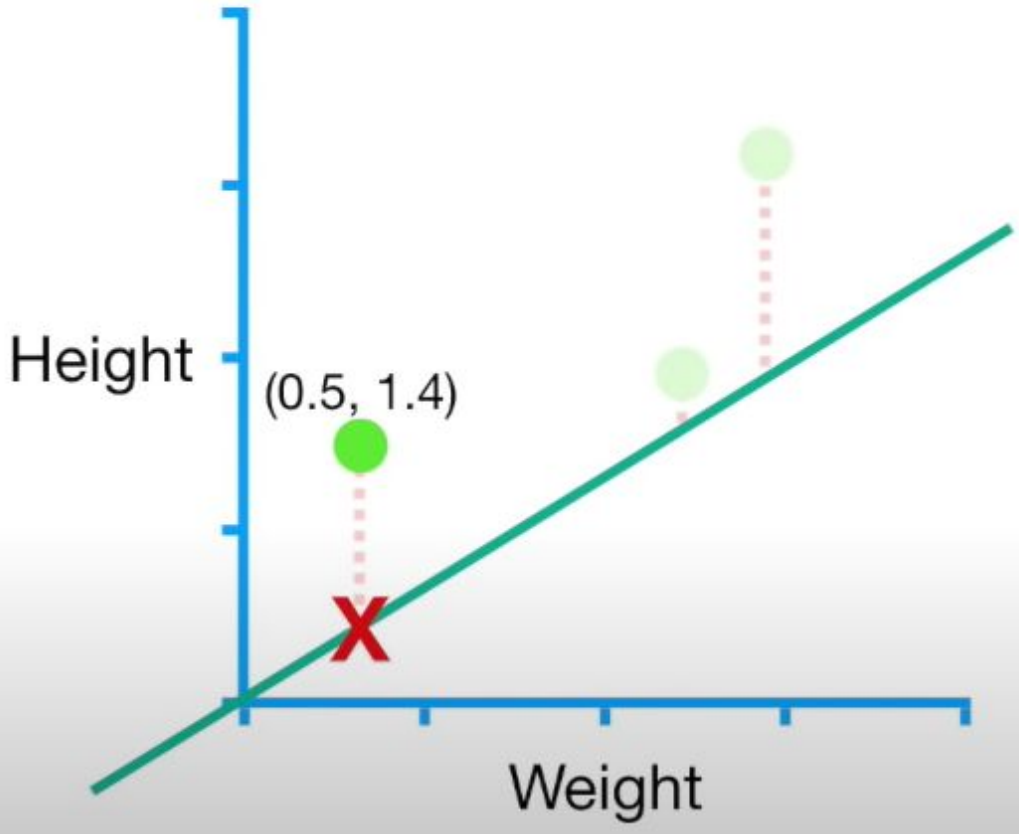



Predicted Height = 0 + 0.64 × Weight
Height
Weight
And that gives us the equation for this line.



In this example, we will evaluate how well this line fits the data with the **Sum of the Squared Residuals**.

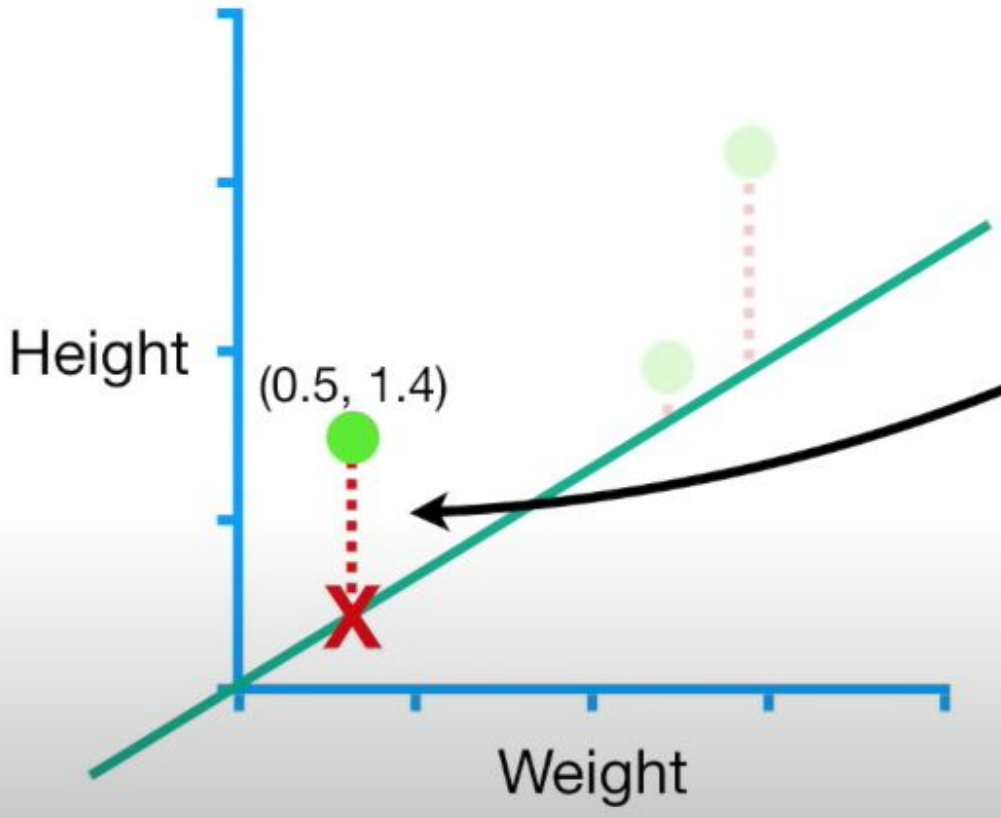
NOTE: In Machine Learning lingo, The Sum of the Squared Residuals is a type of **Loss Function**.



We get the **Predicted Height**, the point on the line…

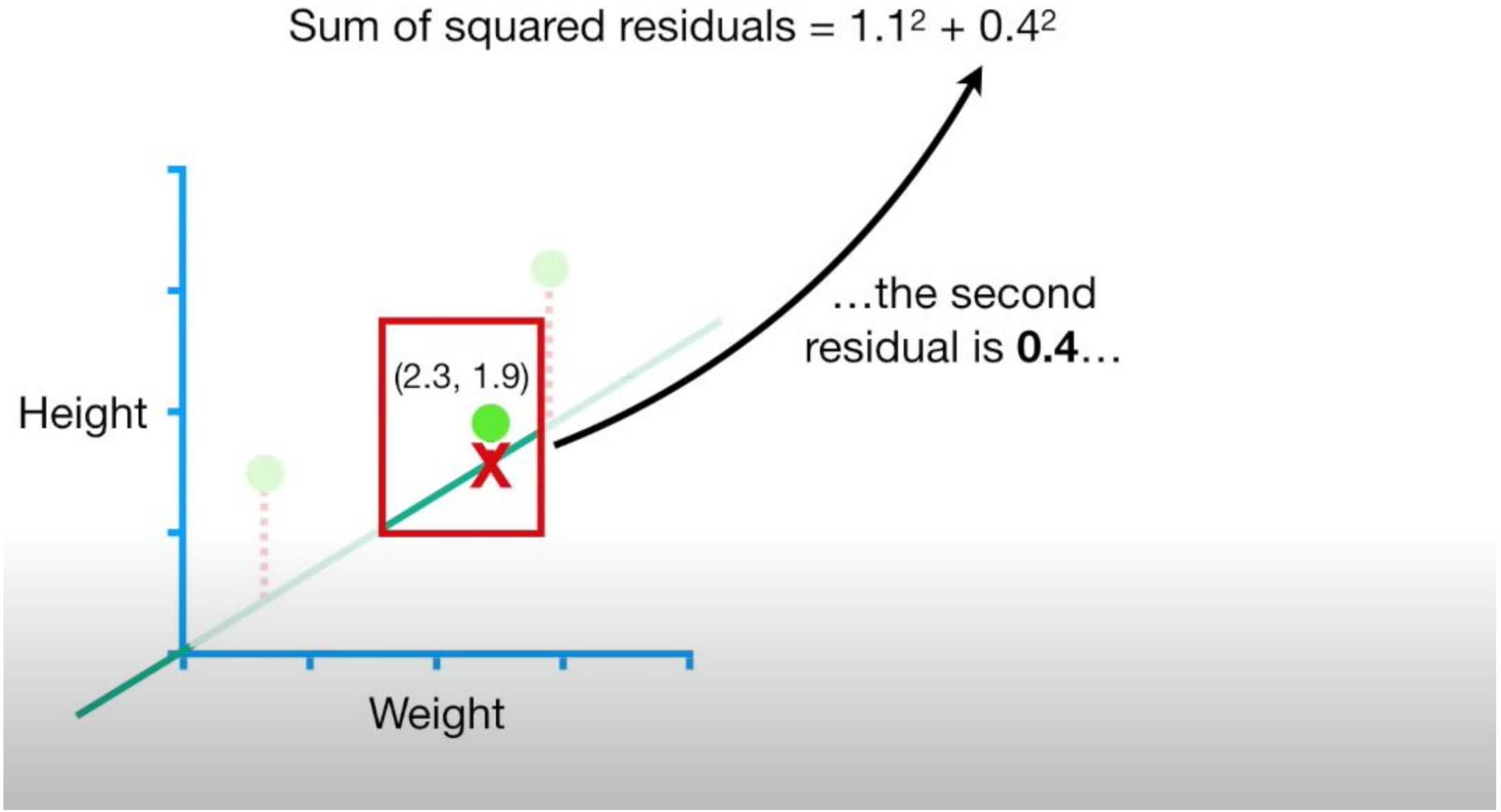
…by plugging **Weight = 0.5** into the equation for the line…

Predicted Height = 0 + 0.64 × **0.5**

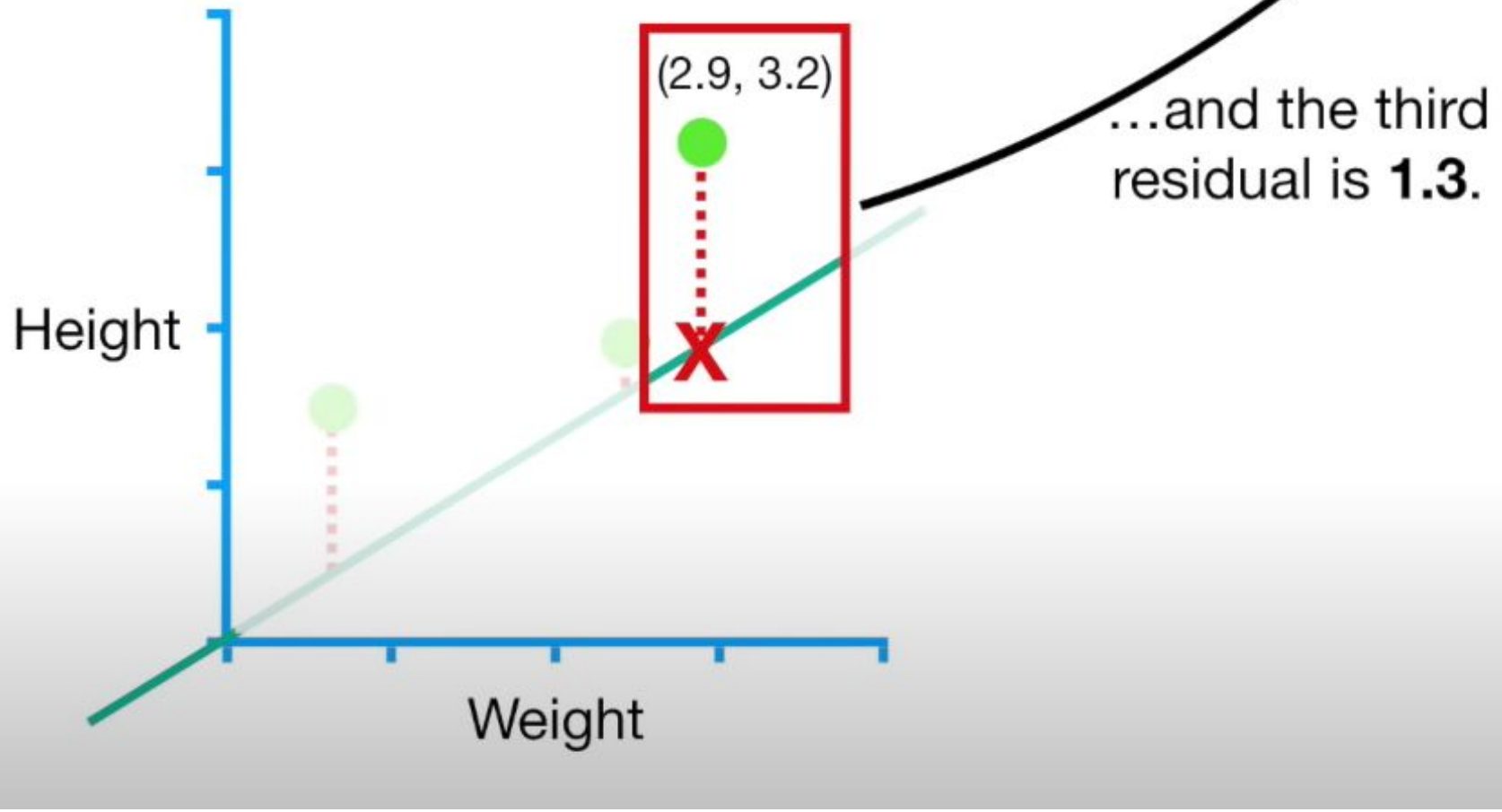


The residual is the difference between the **Observed Height**, and the **Predicted Height**…

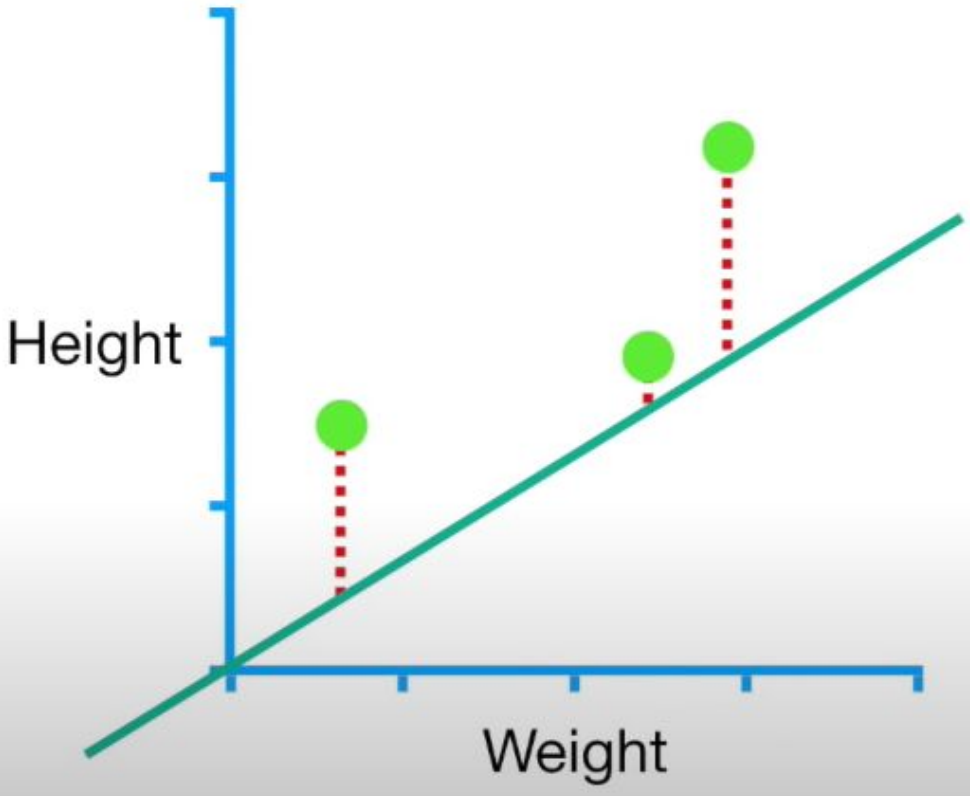
Predicted Height = 0 + 0.64 × **0.5** = **0.32**


Sum of squared residuals = 1.1² + 0.4²
…the second residual is **0.4**…
(2.3, 1.9)
Height
Weight

Sum of squared residuals = $1.1^2 + 0.4^2 + 1.3^2$

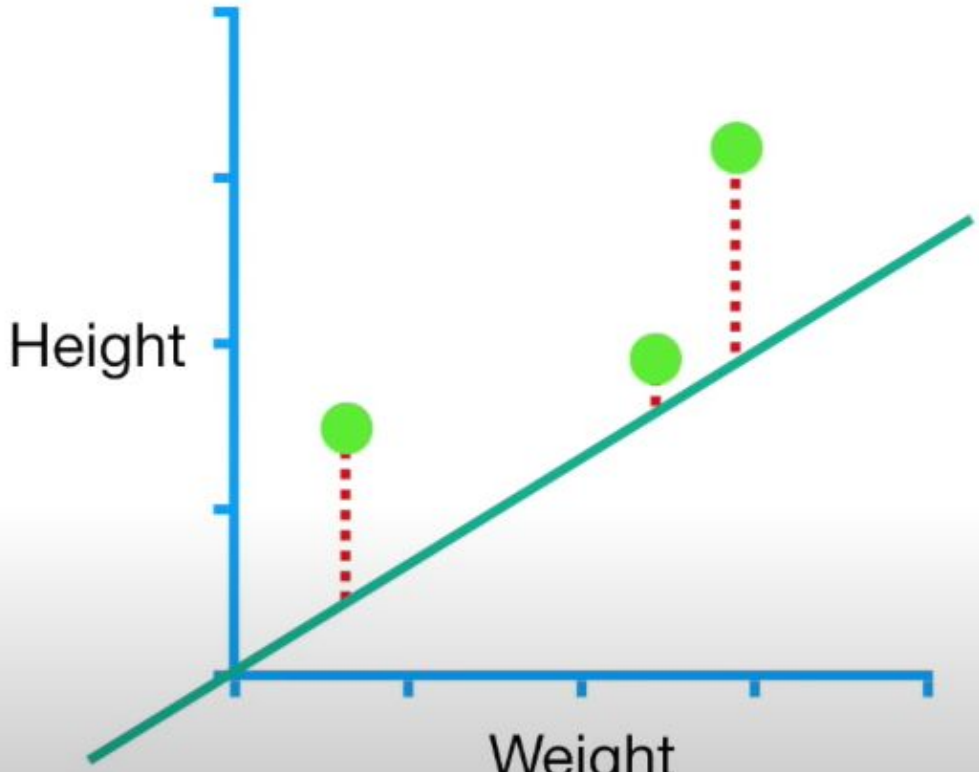


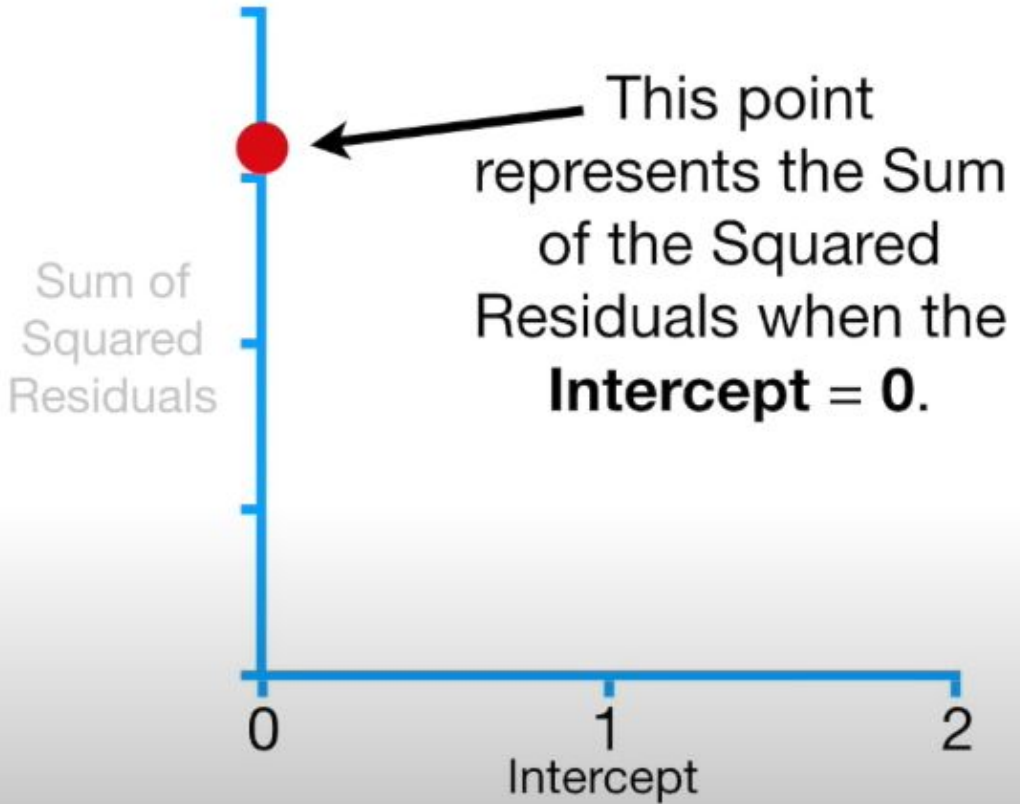
Sum of squared residuals = $1.1^2 + 0.4^2 + 1.3^2 = 3.1$



In the end, **3.1** is the Sum of the Squared Residuals.

Sum of squared residuals = $1.1^2 + 0.4^2 + 1.3^2 = 3.1$

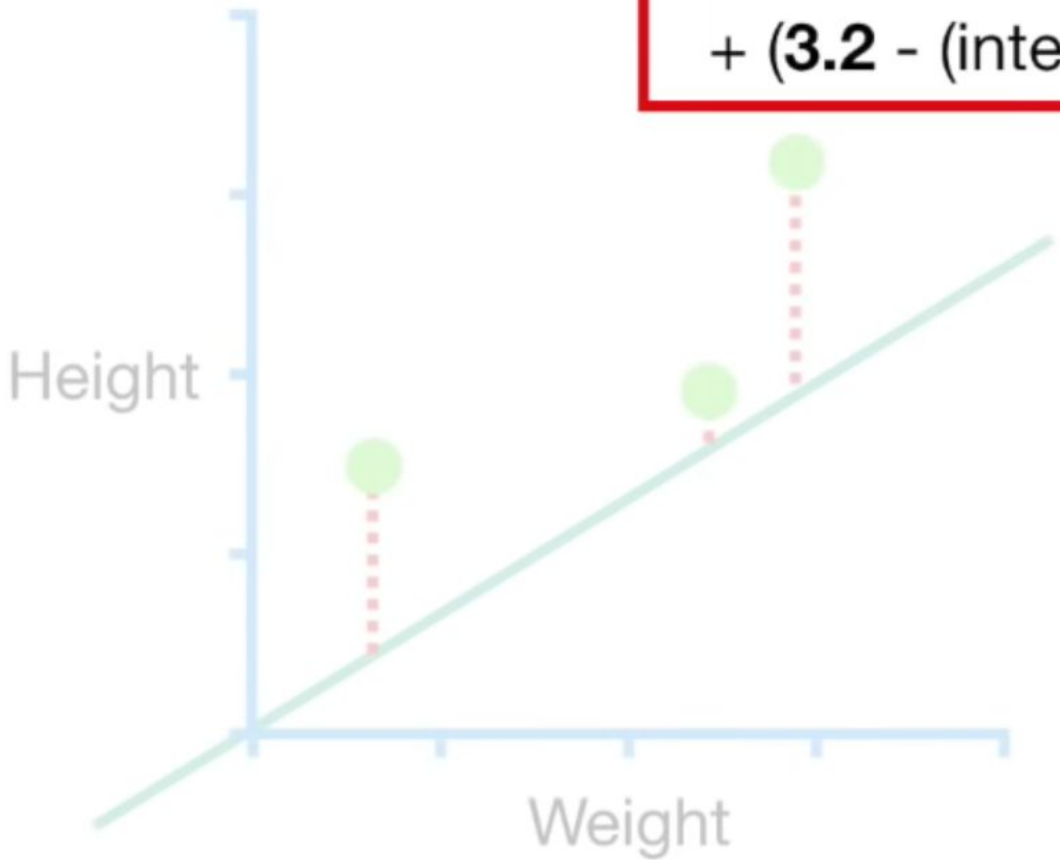




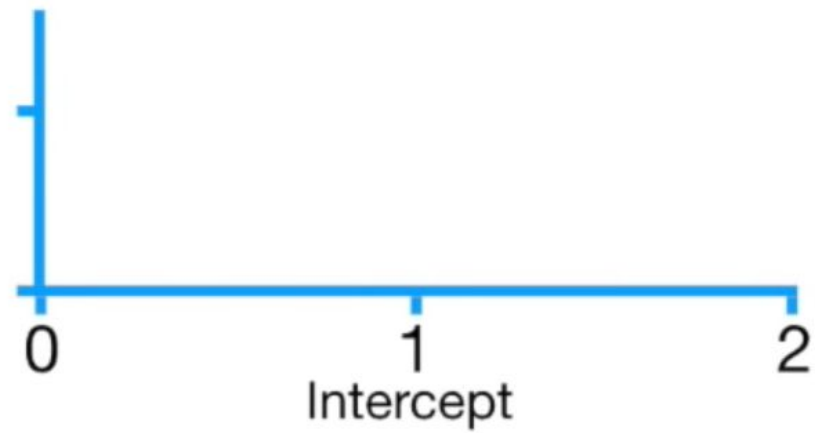
Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

+ (**1.9** - (intercept + 0.64 × **2.3**))^2

+ (**3.2** - (intercept + 0.64 × **2.9**))^2



Now we can easily plug in any value for the **intercept**…

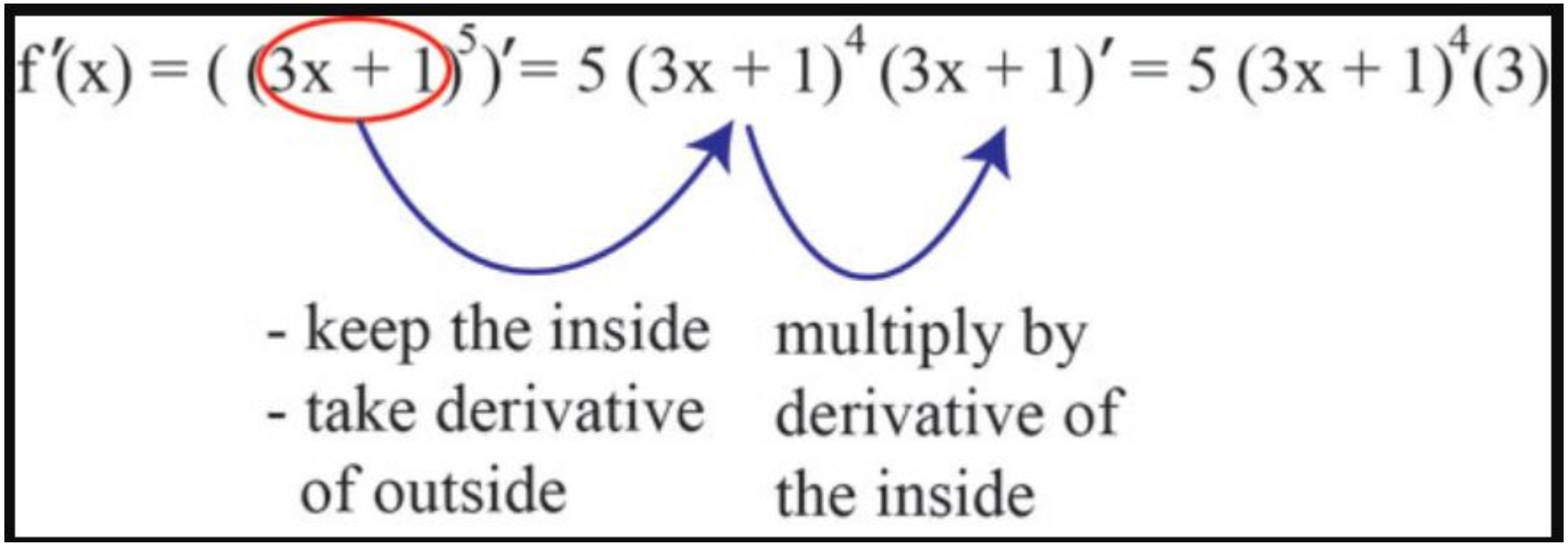


$$\frac{d}{d\ intercept}(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5}))^2 = 2(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5})) \times -1$$

$$\frac{d}{d\ intercept}\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5})$$

$$\frac{d}{d\ intercept}\mathbf{1.4} + (-1)intercept - 0.64 \times \mathbf{0.5} = -1$$

The Chain Rule of differentiation



$$\frac{d}{d\ intercept}(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5}))^2 = 2(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5})) \times -1$$

$$\frac{d}{d\ intercept}\ \mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5})$$

$$\frac{d}{d\ intercept}\ \cancel{\mathbf{1.4}} + (-1)intercept - \cancel{0.64 \times \mathbf{0.5}} = -1$$

These parts don't contain a term for the **Intercept**, so they go away.

Now we need to take the derivative of the next two parts.

$$\frac{d}{d\ intercept}\ \text{Sum of squared residuals} = -2(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5})$$

$$+\ \frac{d}{d\ intercept}\ (\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))^2$$

$$+\ \frac{d}{d\ intercept}\ (\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))^2$$

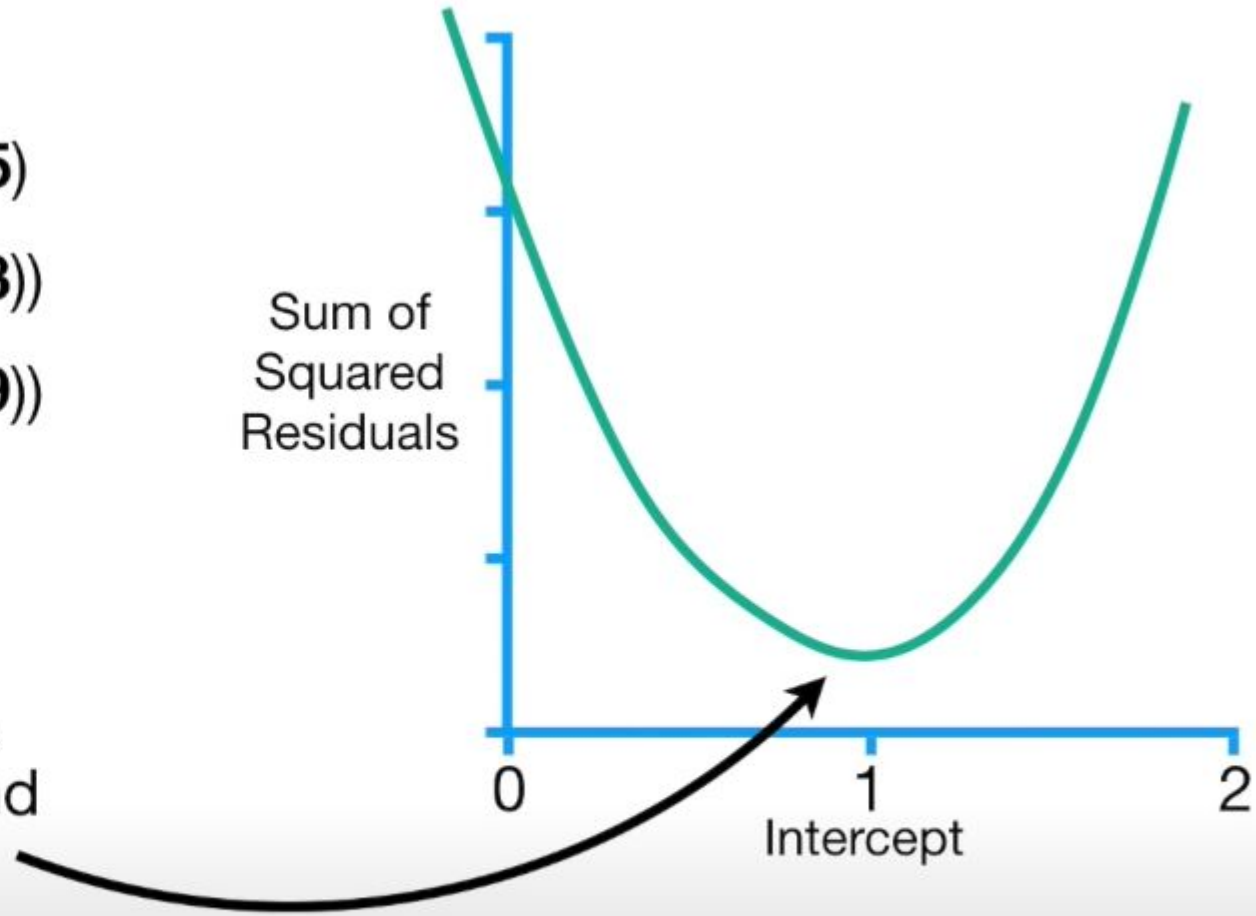
$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

$$-2(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))$$

Now that we have the derivative, **Gradient Descent** will use it to find where the Sum of Squared Residuals is lowest.



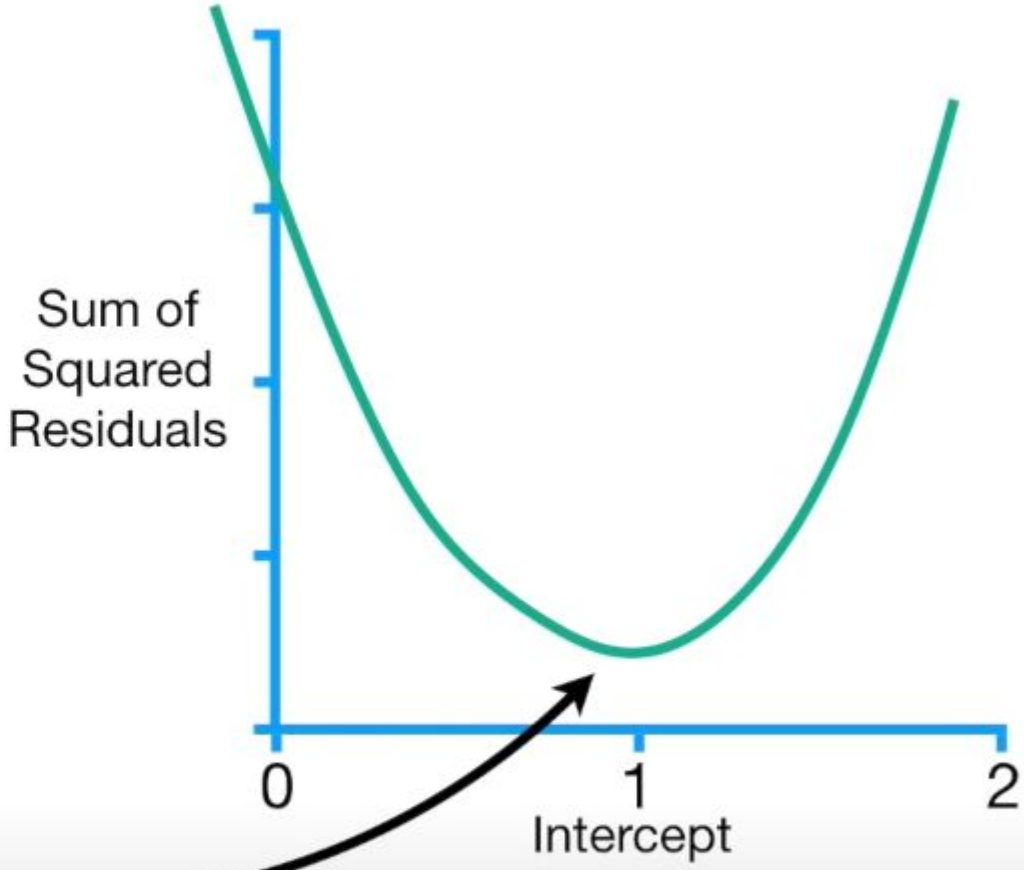
$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

$$-2(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (intercept + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (intercept + 0.64 \times \mathbf{2.9}))$$

NOTE: If we were using **Least Squares** to solve for the optimal value for the **Intercept**, we would simply find where the the slope of the curve = **0**.

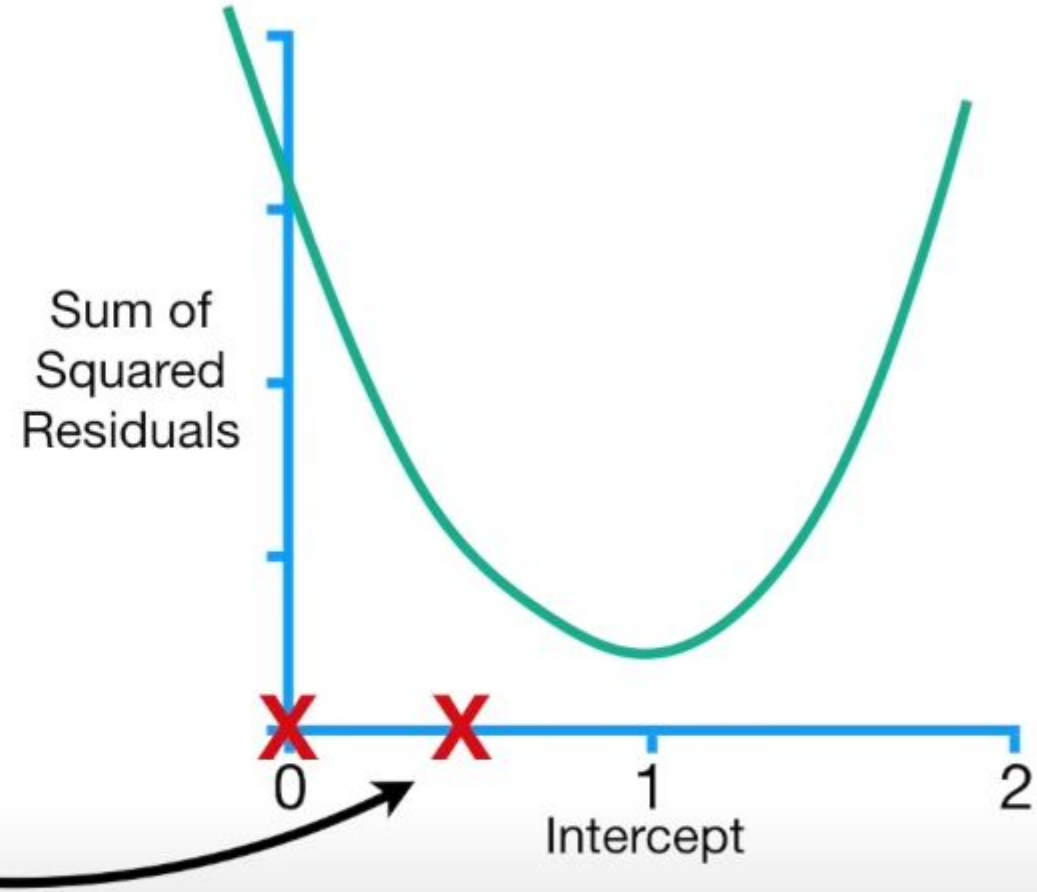


$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

$$-2(\mathbf{1.4} - (intercept + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (intercept + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (intercept + 0.64 \times \mathbf{2.9}))$$



In contrast, **Gradient Descent** finds the minimum value by taking steps from an initial guess until it reaches the best value.

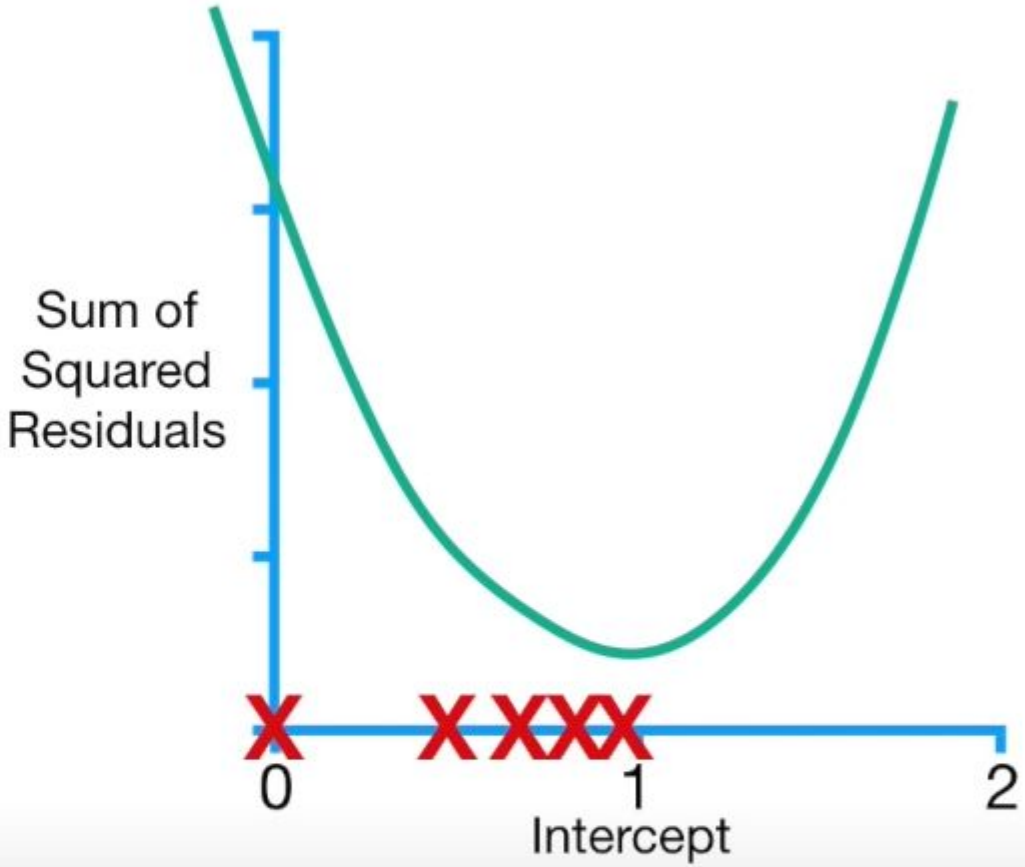
$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

$$-2(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (\text{intercept} + 0.64 \times \mathbf{2.9}))$$

This makes **Gradient Descent** very useful when it is not possible to solve for where the derivative = **0**, and this is why **Gradient Descent** can be used in so many different situations.



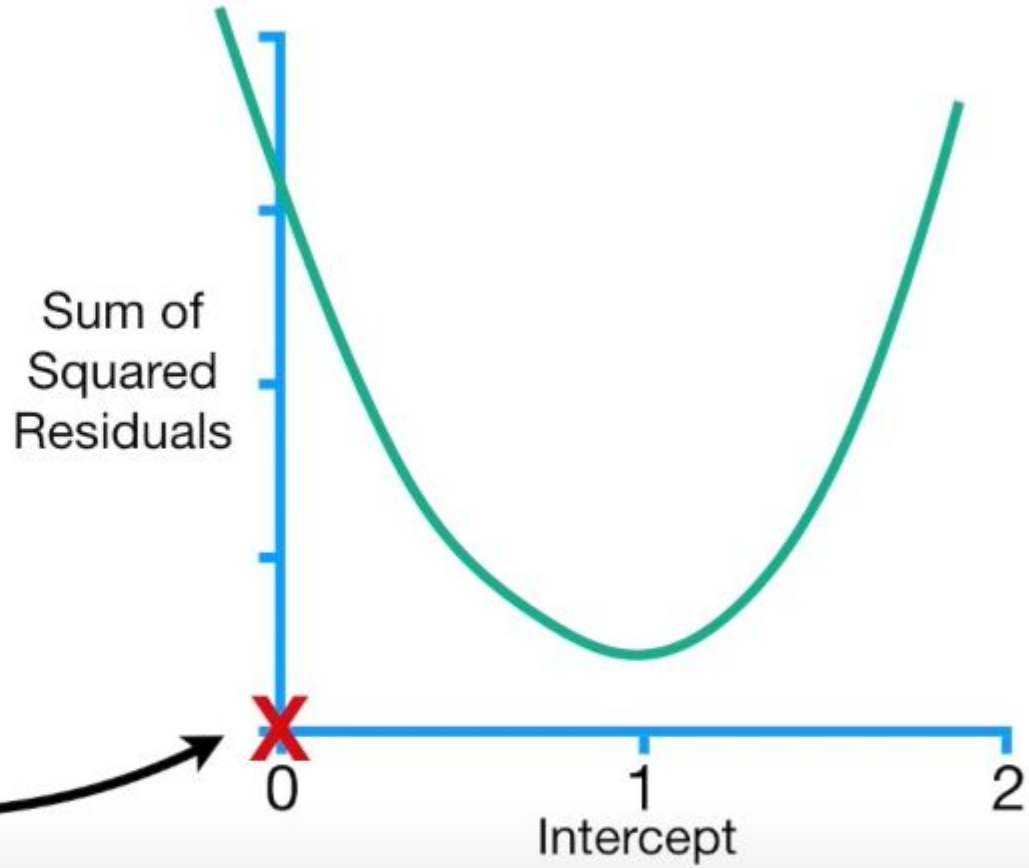
$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

-2(**1.4** - (intercept + 0.64 × **0.5**)

+ -2(**1.9** - (intercept + 0.64 × **2.3**))

+ -2(**3.2** - (intercept + 0.64 × **2.9**))

Remember, we started by setting the **Intercept** to a random number. In this case, that was **0**.



$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

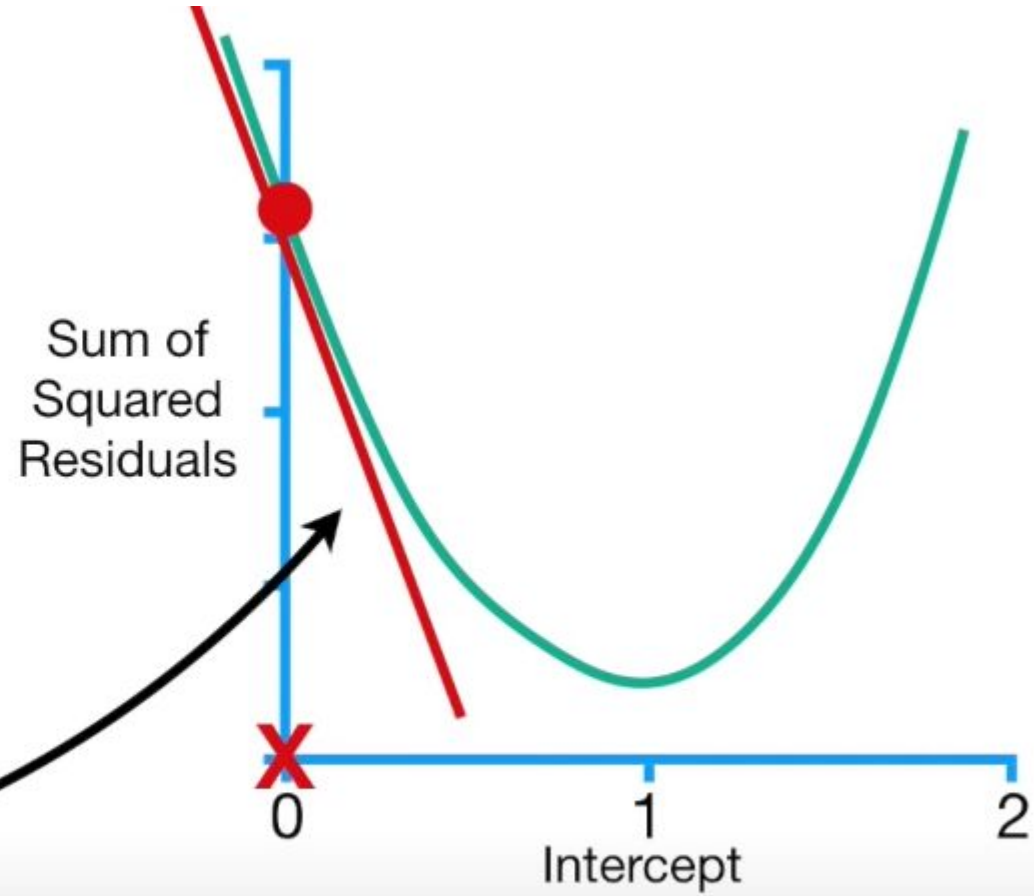
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$

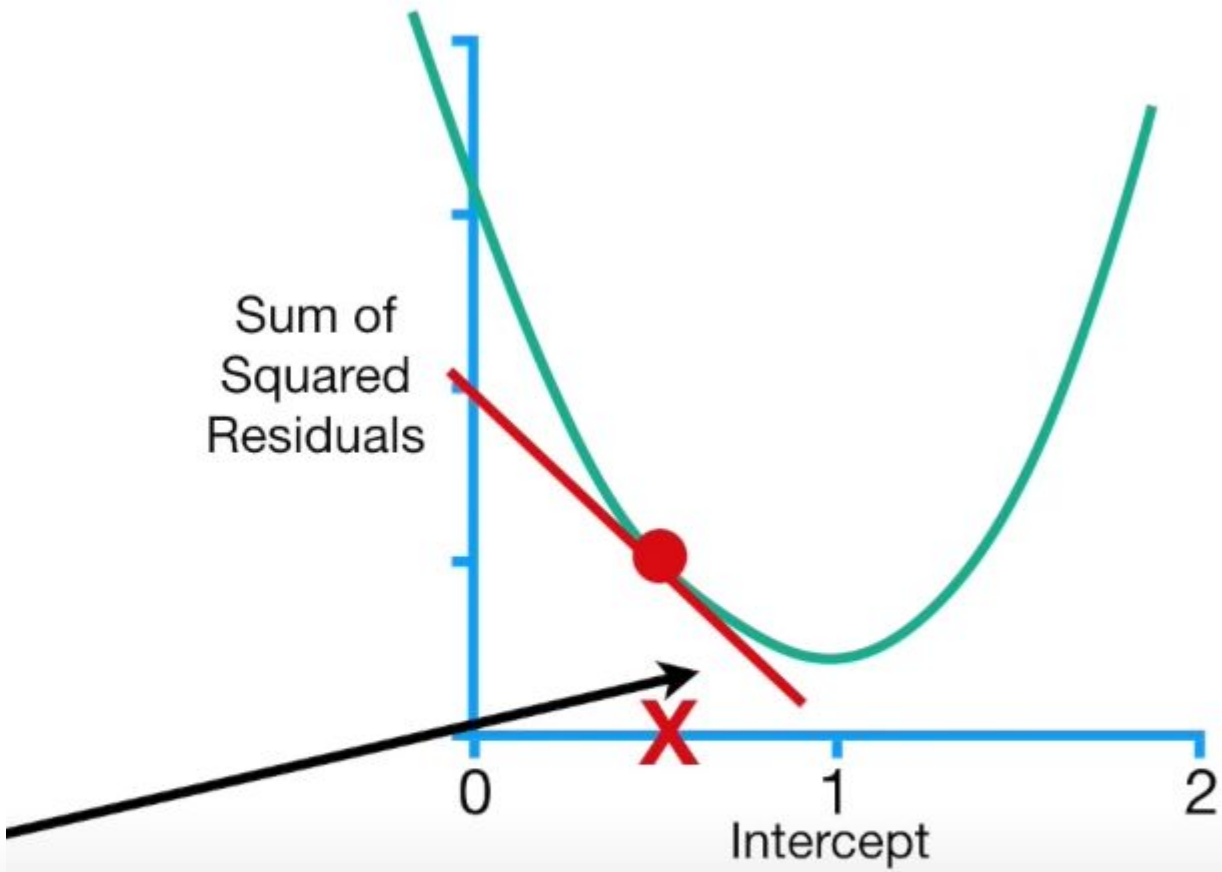
So when the **Intercept** = **0**,
the slope of the curve = **-5.7**.



NOTE: The closer we get to the optimal value for the **Intercept**, the closer the slope of the curve gets to **0**.

This means that when the slope of the curve is close to **0**…

…then we should take baby steps, because we are close to the optimal value…



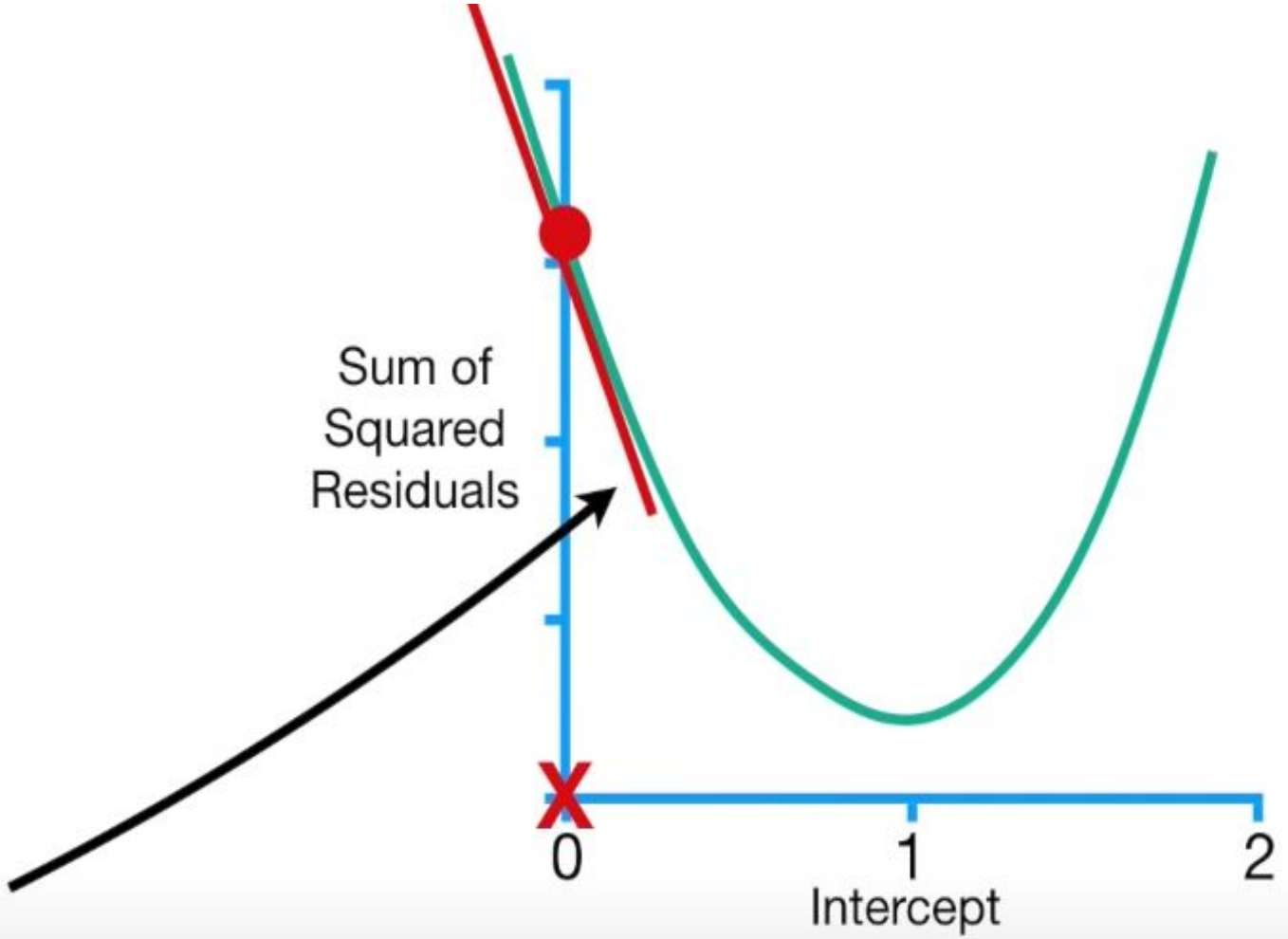
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$



…and when the slope is far from **0**…

…then we should take big steps, because we are far from the optimal value.

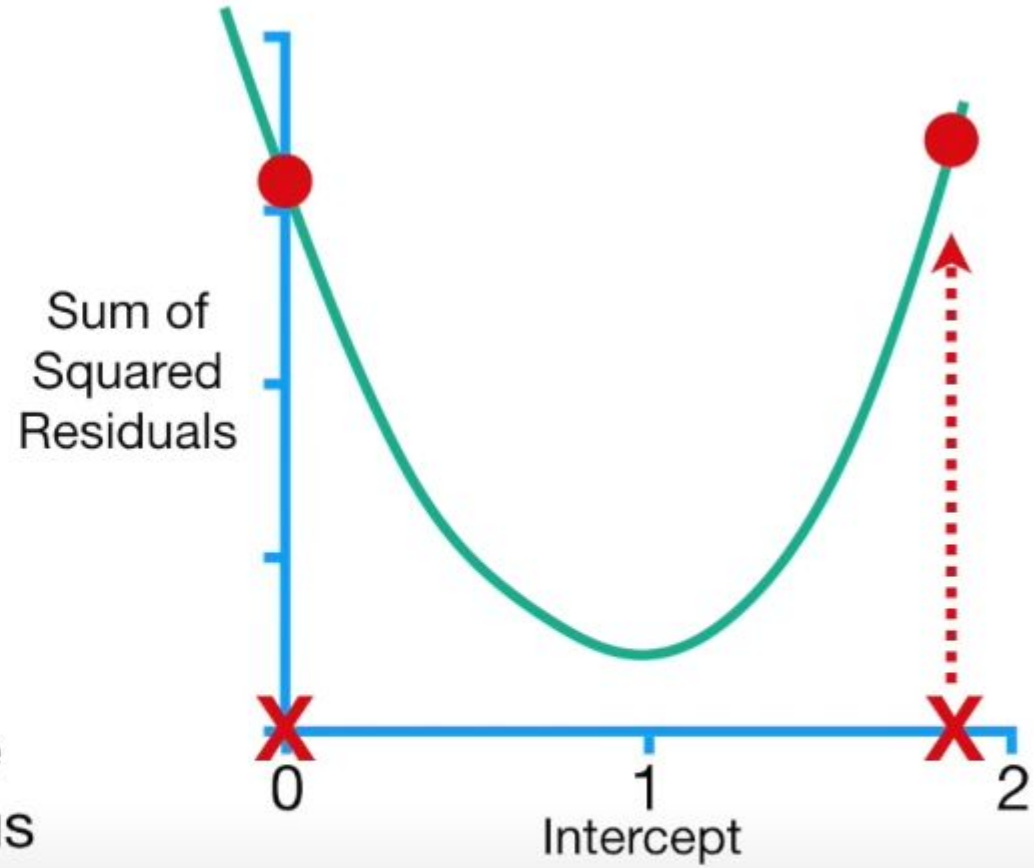
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$



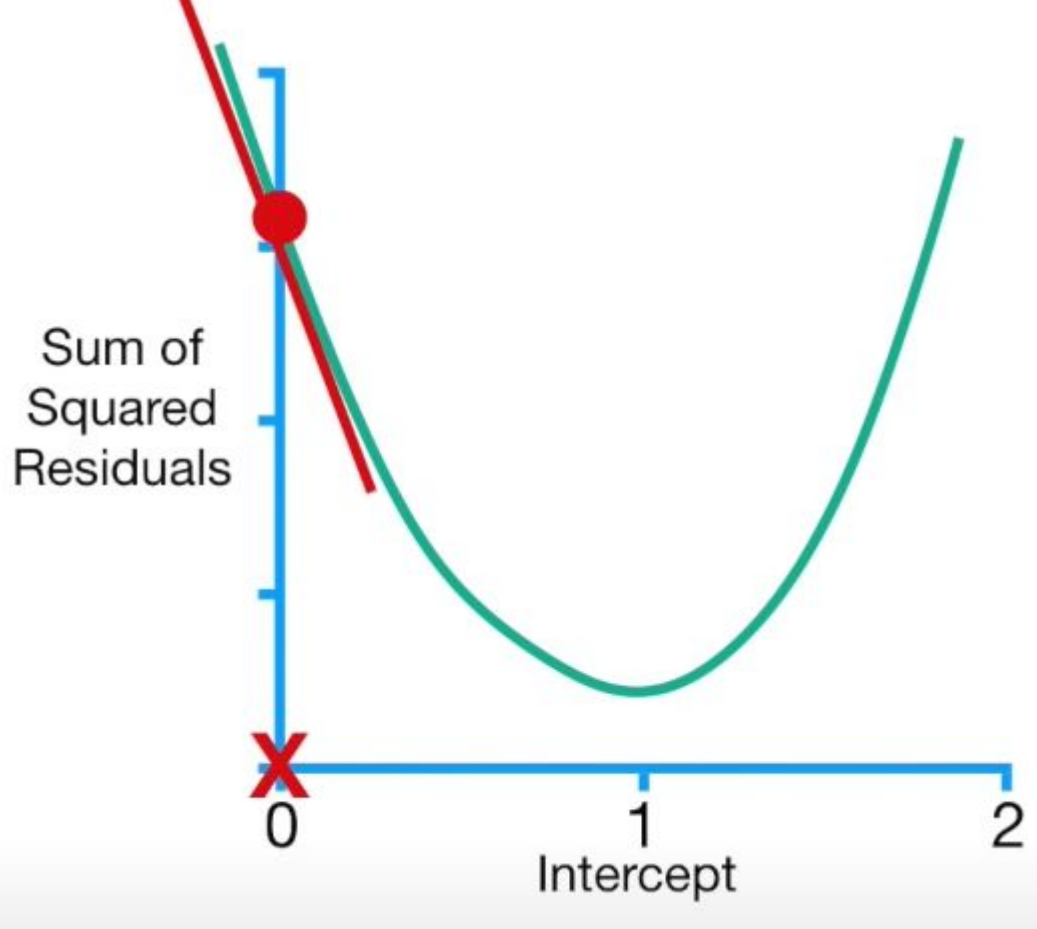
So the size of the step should be related to the slope, since it tells us if we should take a baby step or a big step, but we need to make sure the big step is not too big.

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} = -2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5}) + -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3})) + -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$

Step Size = -5.7 x 0.1

Gradient Descent determines the **Step Size** by multiplying the **slope**……by a small number called **The Learning Rate**.



$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

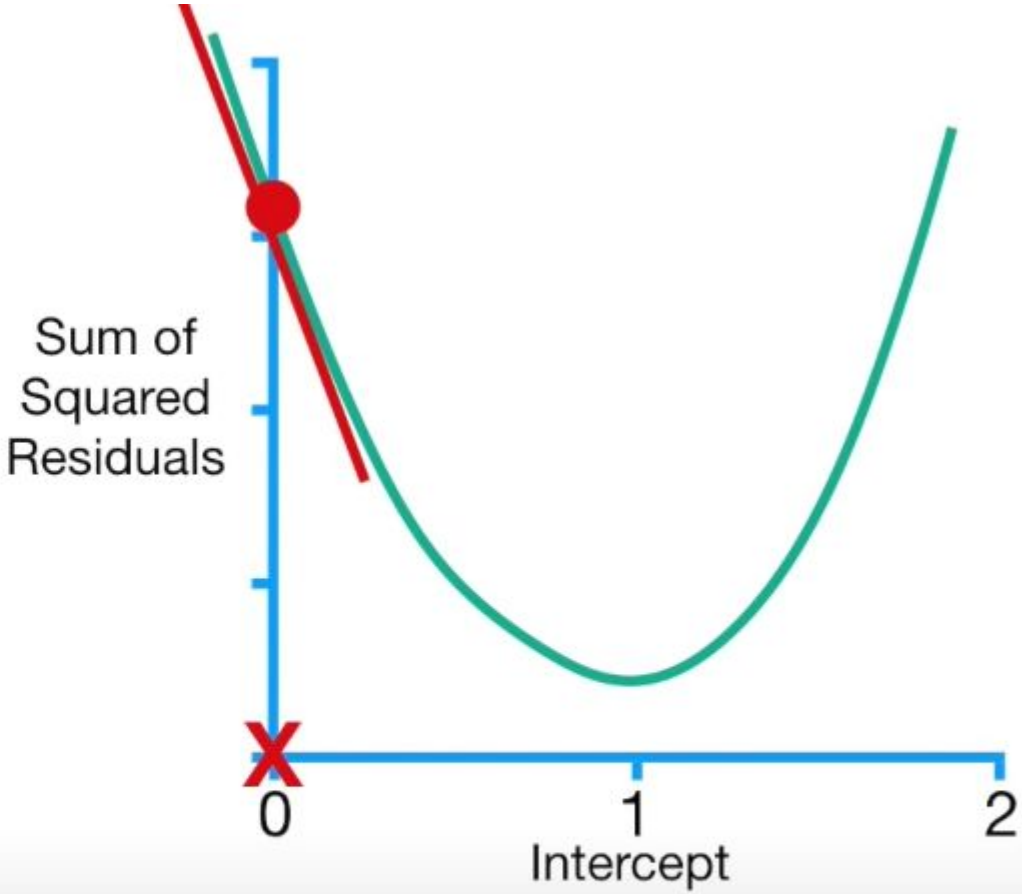
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5}))$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$

Step Size = -5.7 x 0.1 = -0.57



When the **Intercept** = **0**, the **Step Size** = **-0.57**.

New Intercept = **Old Intercept** - **Step Size**

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

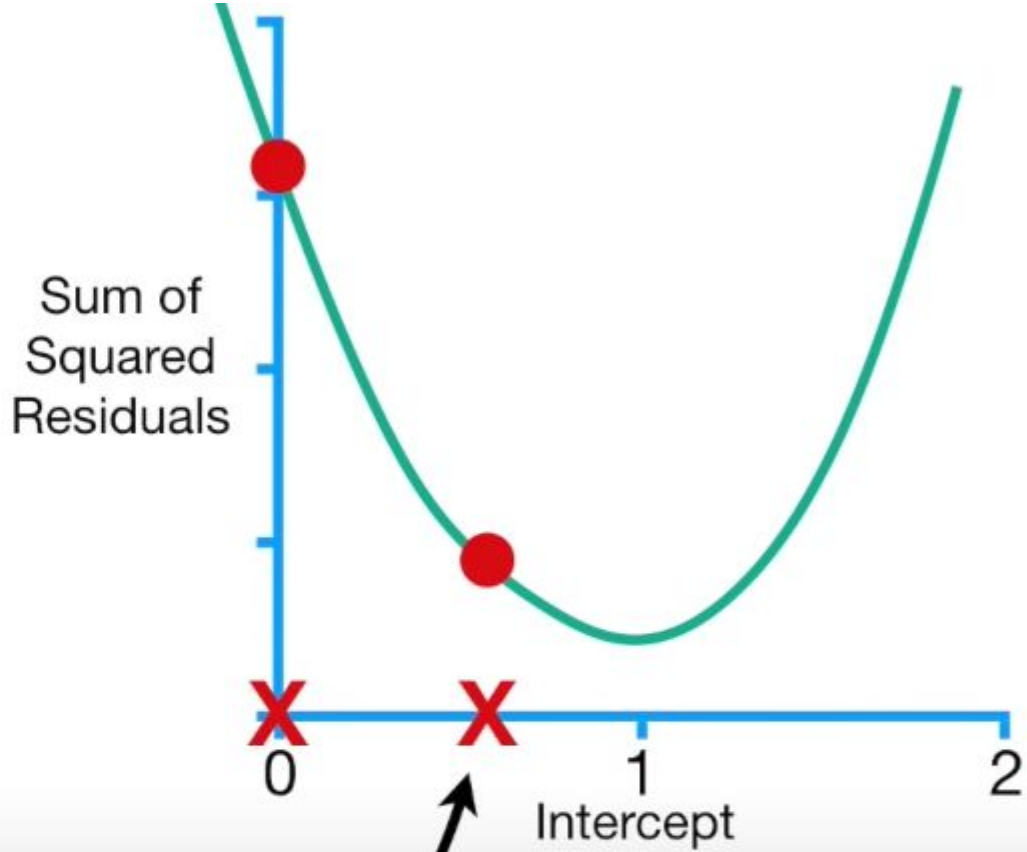
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

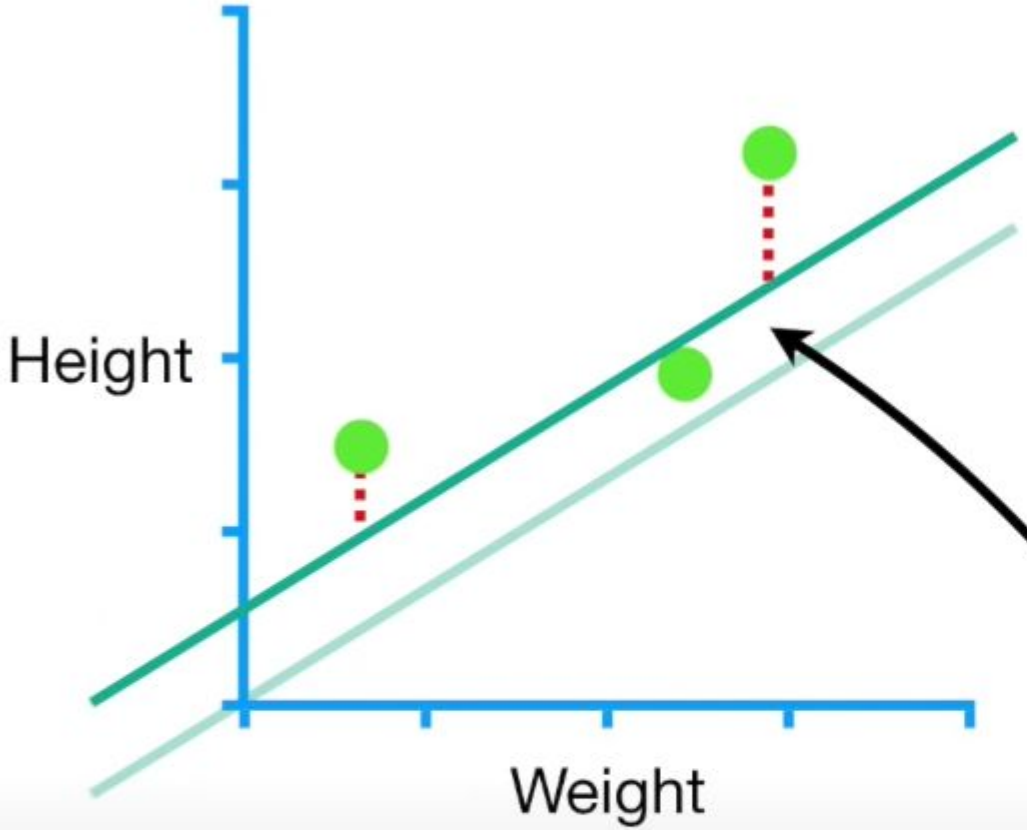
$$= -5.7$$

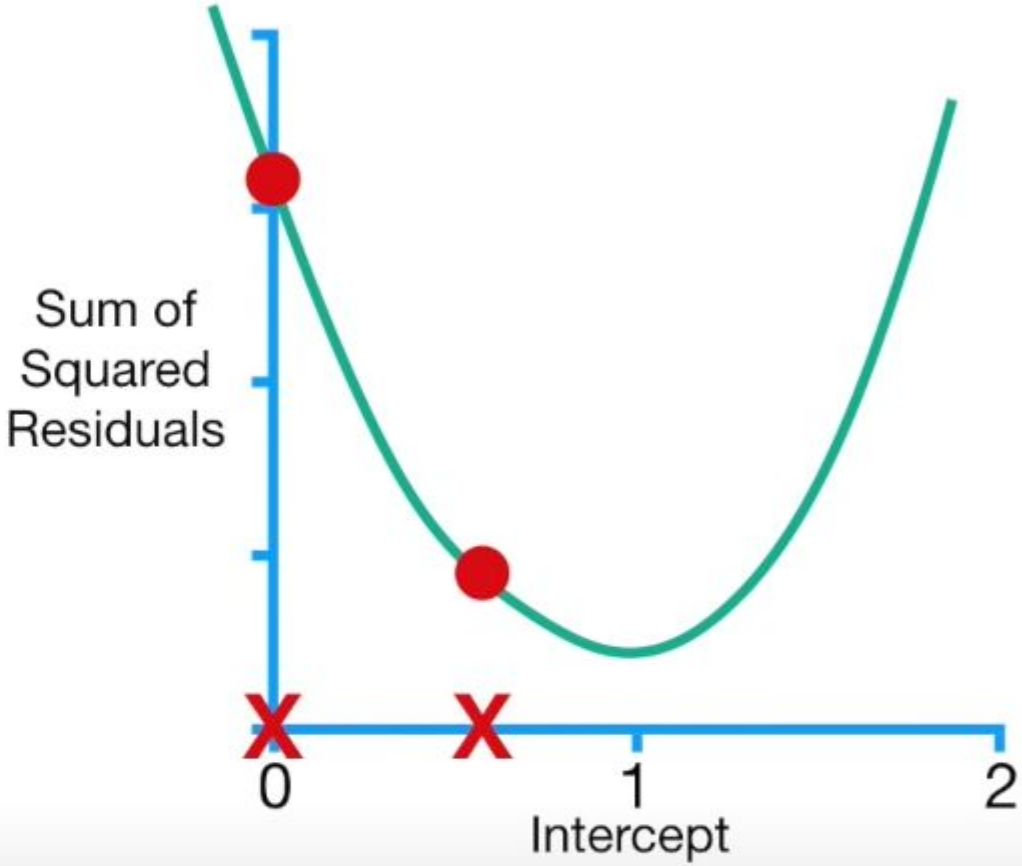
Step Size $= -5.7 \times 0.1 = -0.57$



New Intercept $= 0 - (-0.57) =$ **0.57**

…and the the **New Intercept** = **0.57**.





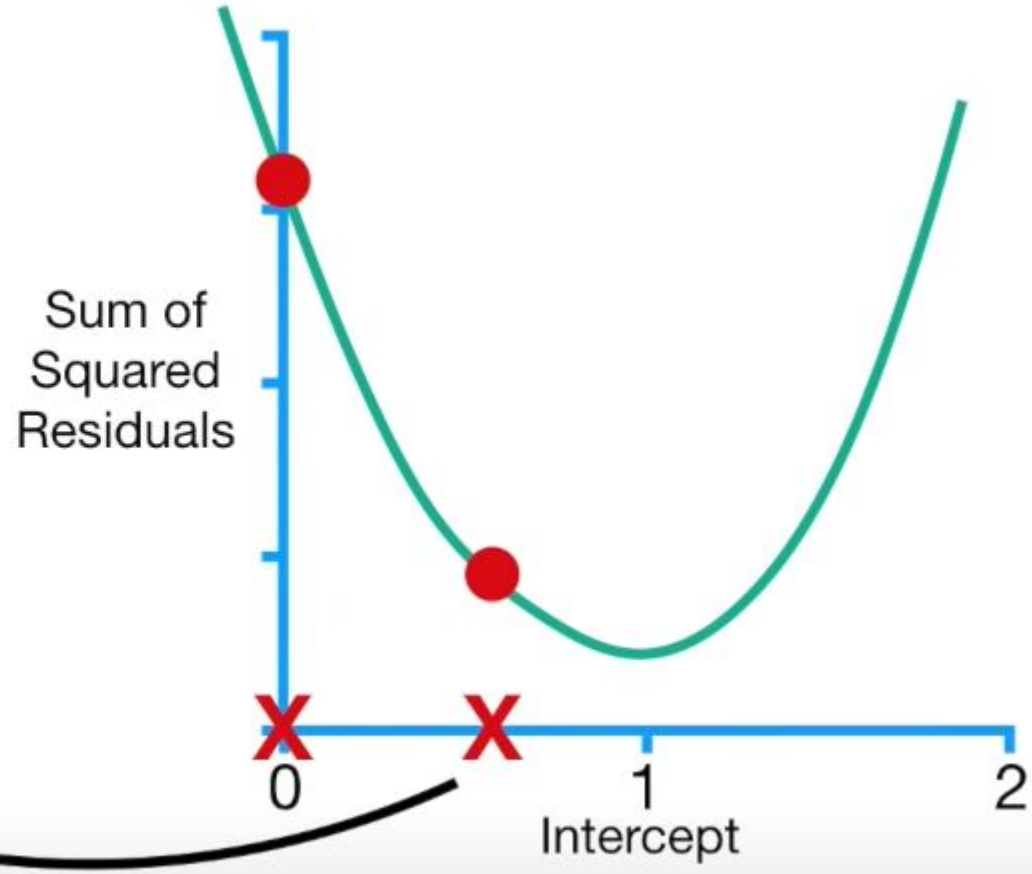
…we can see how much the residuals shrink when the **Intercept = 0.57**.

$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

-2(**1.4** - (intercept + 0.64 × **0.5**)

+ -2(**1.9** - (intercept + 0.64 × **2.3**))

+ -2(**3.2** - (intercept + 0.64 × **2.9**))



To take another step, we go back to the derivative and plug in the **New Intercept** (**0.57**)…

$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

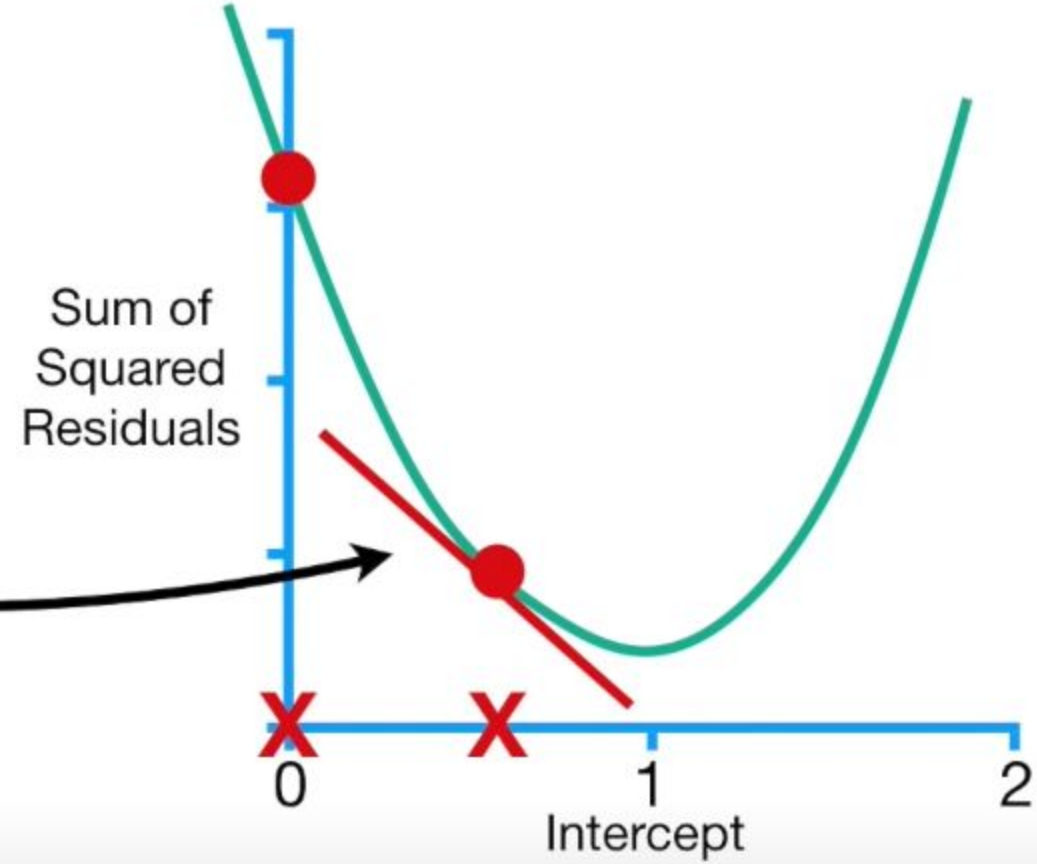
$$-2(\mathbf{1.4} - (0.57 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0.57 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0.57 + 0.64 \times \mathbf{2.9}))$$

$$= \boxed{-2.3}$$

…and that tells us the slope of the curve = **-2.3**.



Step Size = **Slope** × **Learning Rate**

Now let's calculate the **Step Size**…

$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

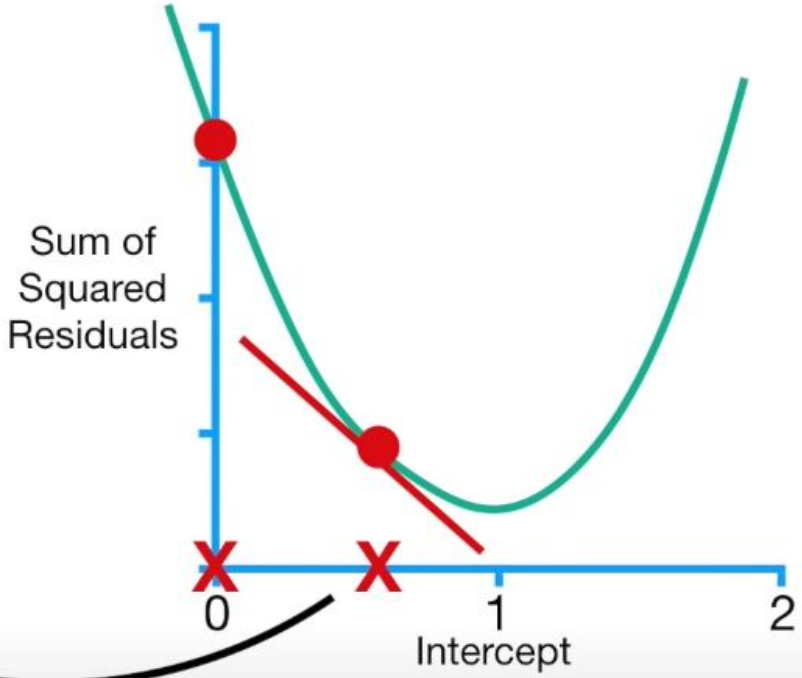
$$-2(\mathbf{1.4} - (0.57 + 0.64 \times \mathbf{0.5}))$$

$$+ -2(\mathbf{1.9} - (0.57 + 0.64 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0.57 + 0.64 \times \mathbf{2.9}))$$

$$= -2.3$$

Step Size = -2.3 x 0.1 = **-0.23**

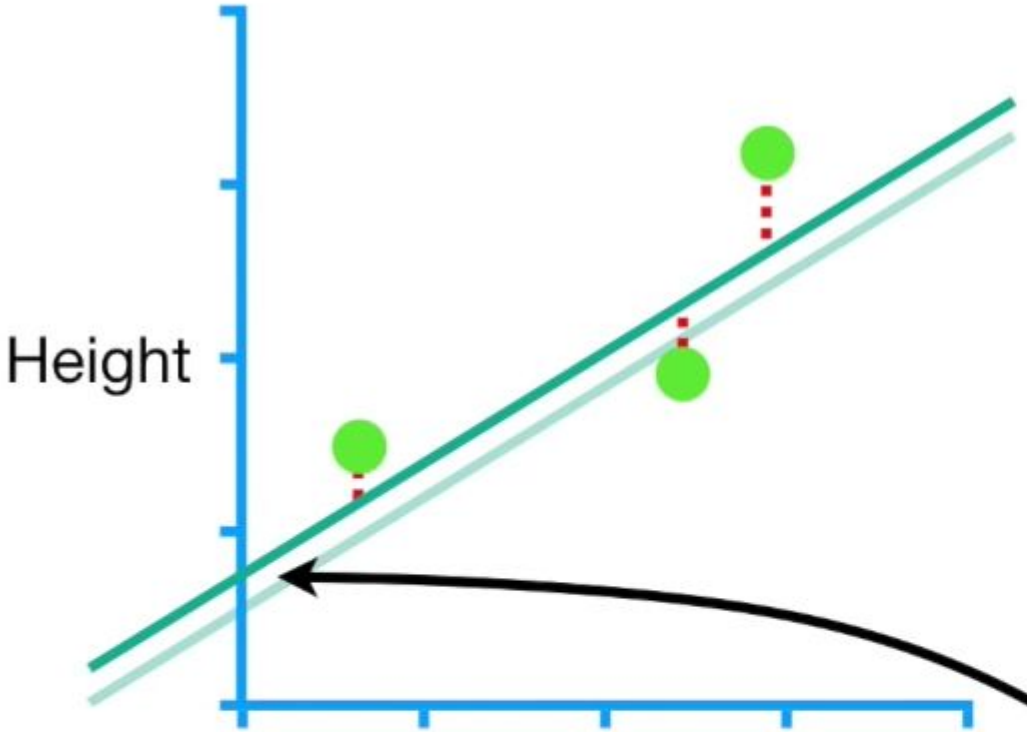


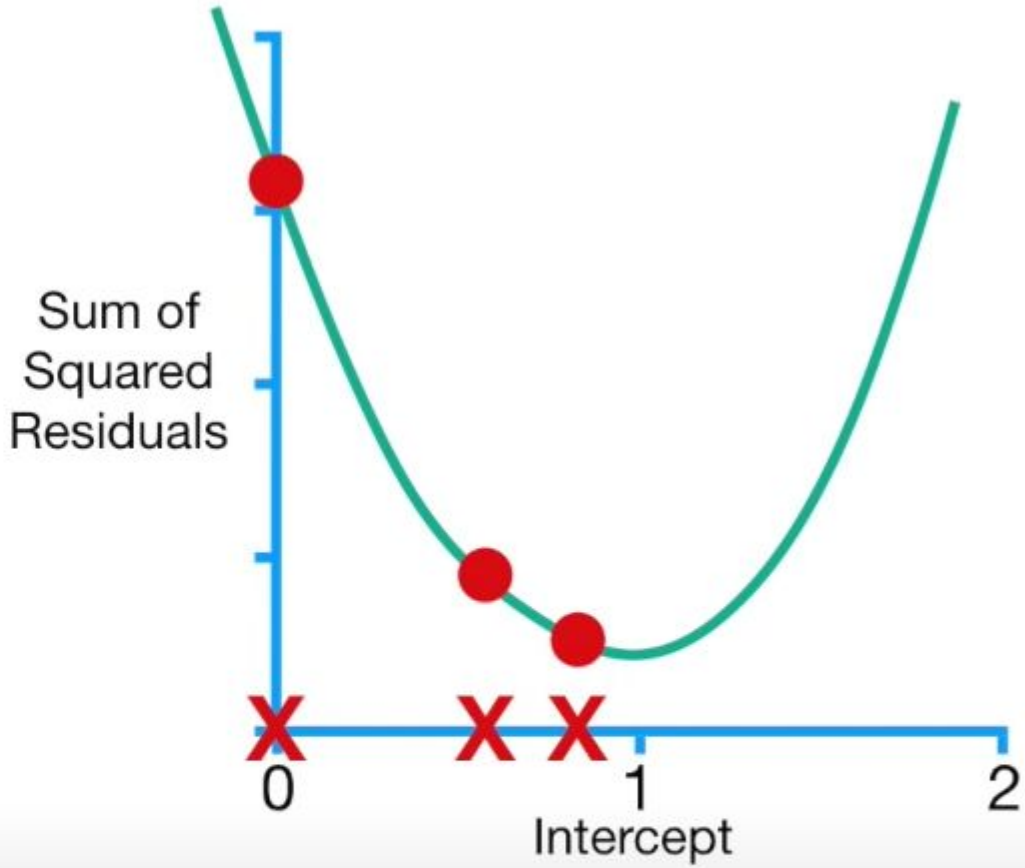
New Intercept = 0.57 - **Step Size**

…and the **New Intercept**…

New Intercept = 0.57 - (-0.23) = **0.8**

…and the **New Intercept** = **0.8**





…to when the
Intercept = 0.8

$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

$$-2(\mathbf{1.4} - (0.8 + 0.64 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0.8 + 0.64 \times \mathbf{2.3}))$$

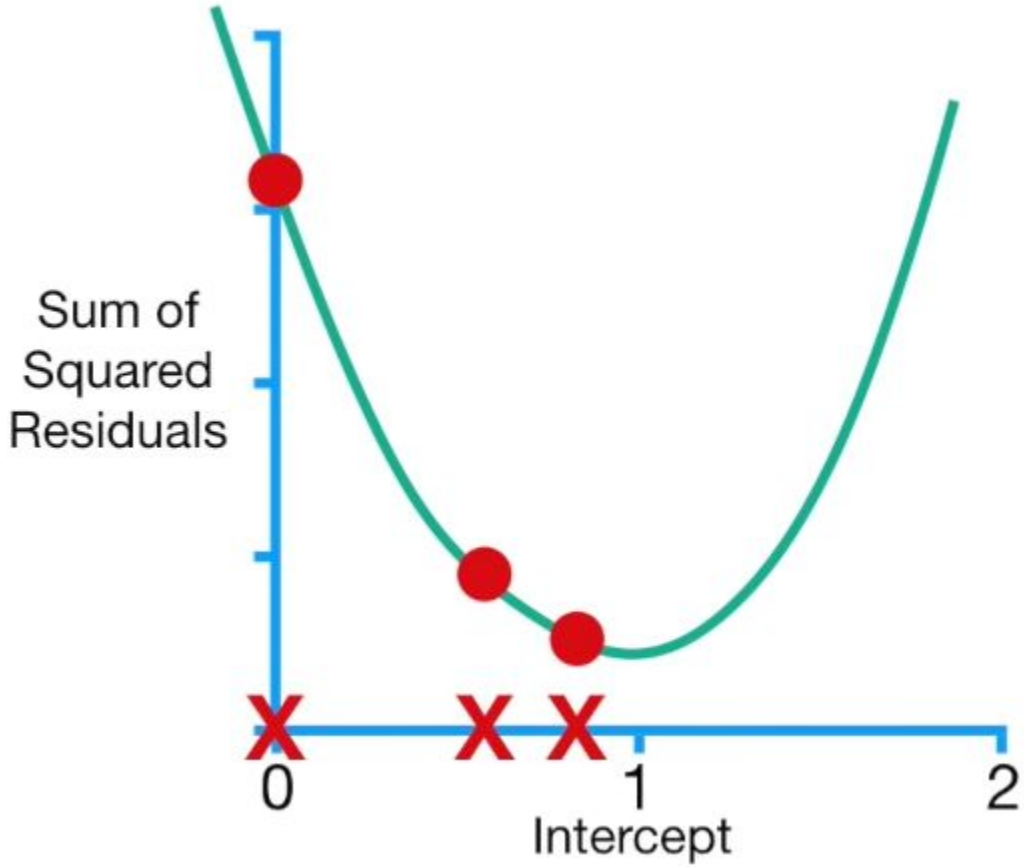
$$+ -2(\mathbf{3.2} - (0.8 + 0.64 \times \mathbf{2.9}))$$

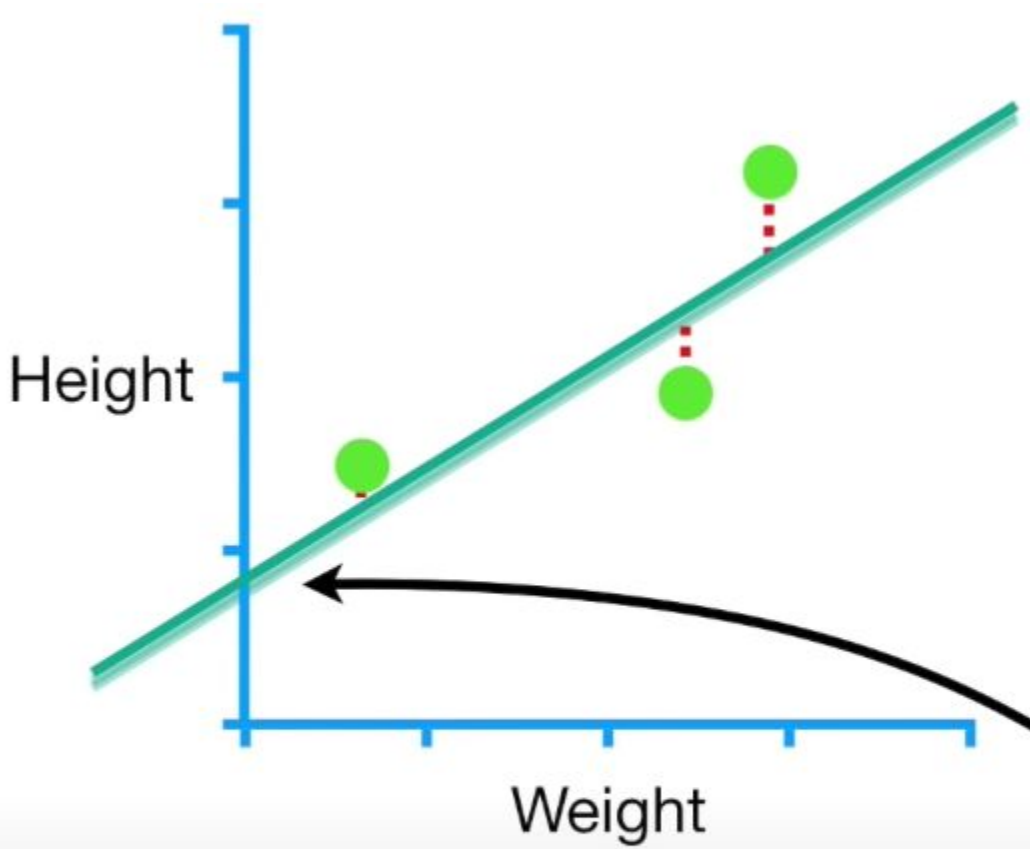
$$= \boxed{\mathbf{-0.9}}$$

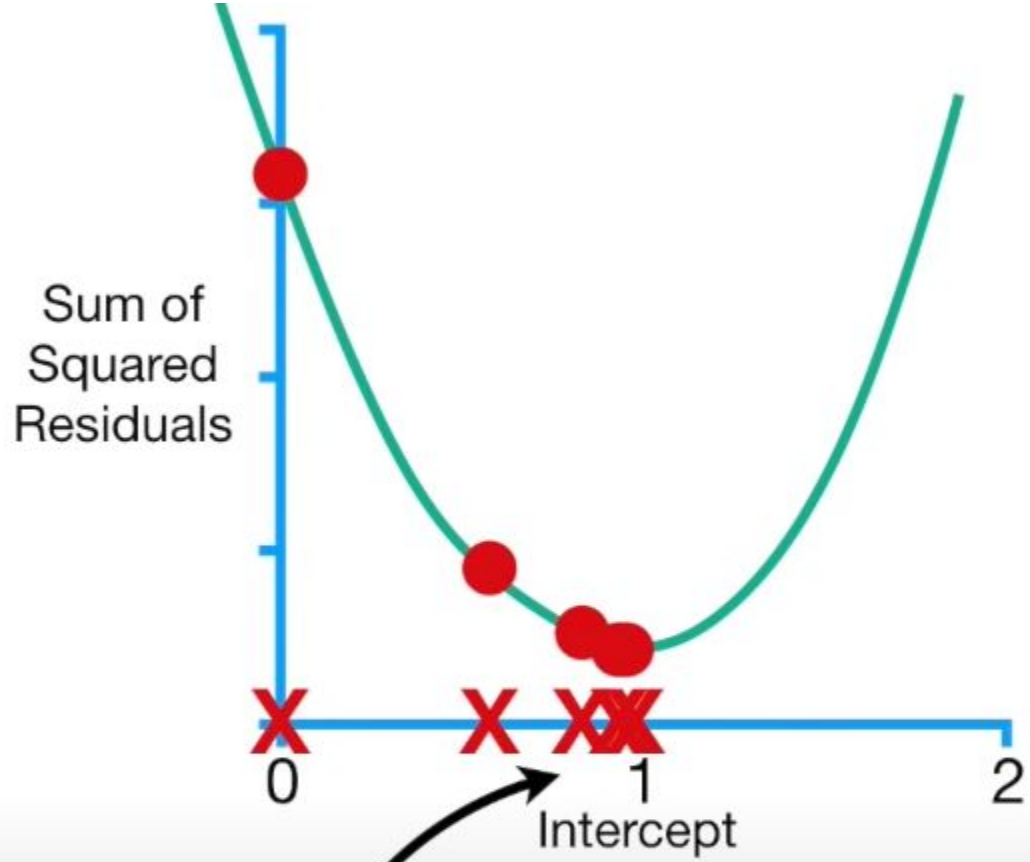
…and we get **-0.9**.

Step Size = -0.9 × 0.1 = **-0.09**

New Intercept = 0.8 - (-0.09) = **0.89**







Then we take another step and the **New Intercept** = **0.92**…

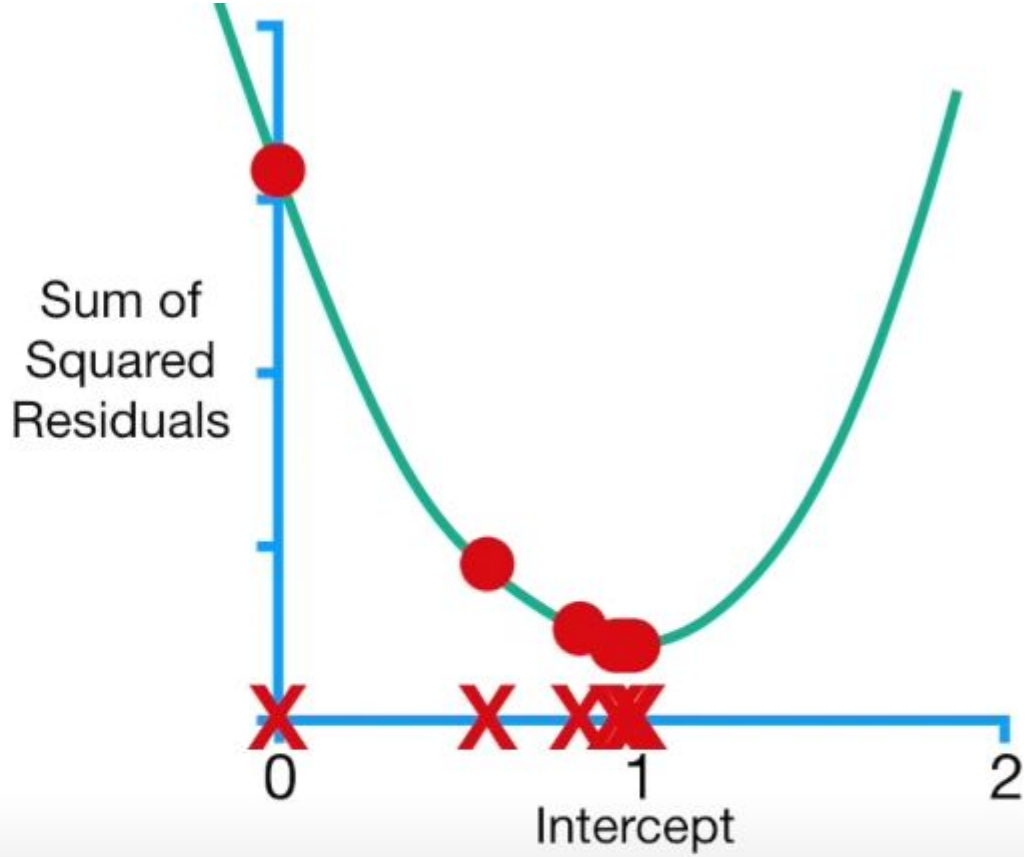
…and then we take another step and the **New Intercept** = **0.94**…

…and then we take another step and the **New Intercept** = **0.95**.

After 6 steps, the **Gradient Descent** estimate for the **Intercept** is **0.95**.

NOTE: The **Least Squares** estimate for the intercept is also **0.95**.

So we know that **Gradient Descent** has done its job, but without comparing its solution to a gold standard, how does **Gradient Descent** know to stop taking steps?

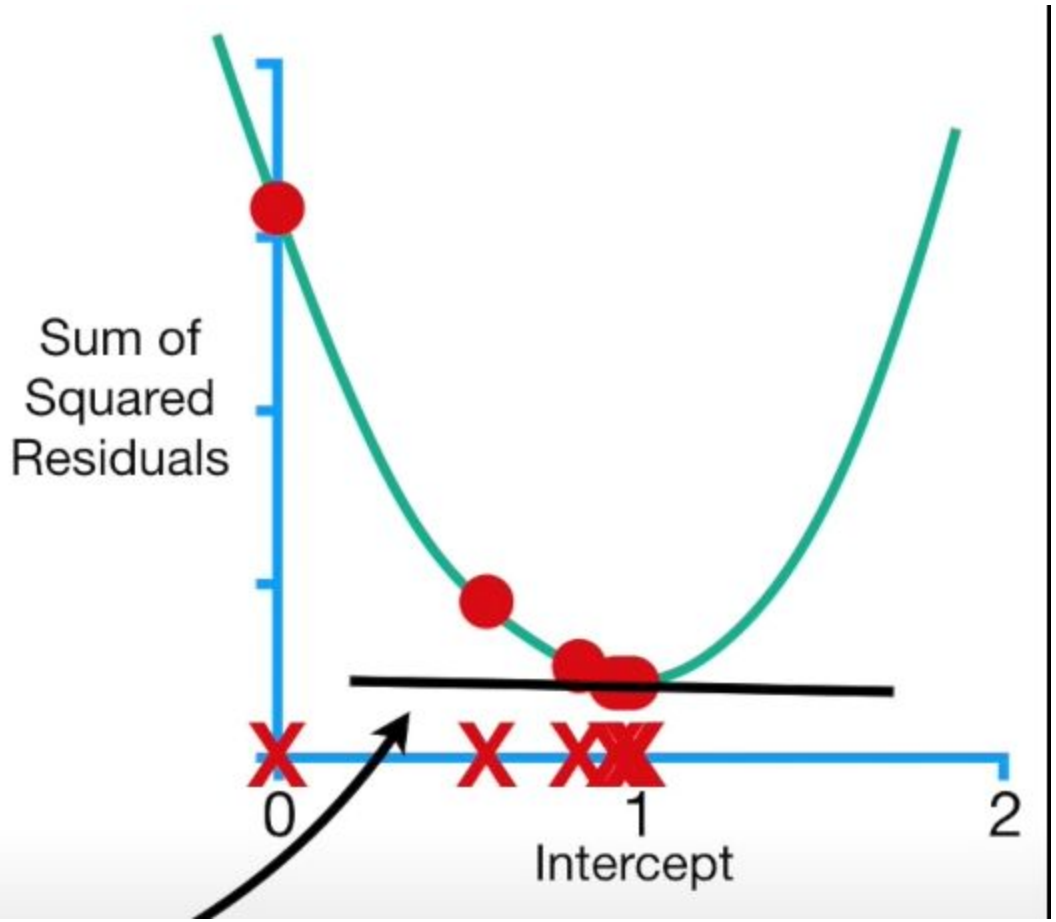


Gradient Descent stops when the **Step Size** is **Very Close To 0**.

Step Size = **Slope** × **Learning Rate**

The **Step Size** will be **Very Close to 0** when the **Slope** is very close to **0**.

Step Size = **Slope** × **Learning Rate**



In practice, the
Minimum Step Size = **0.001**
or smaller.

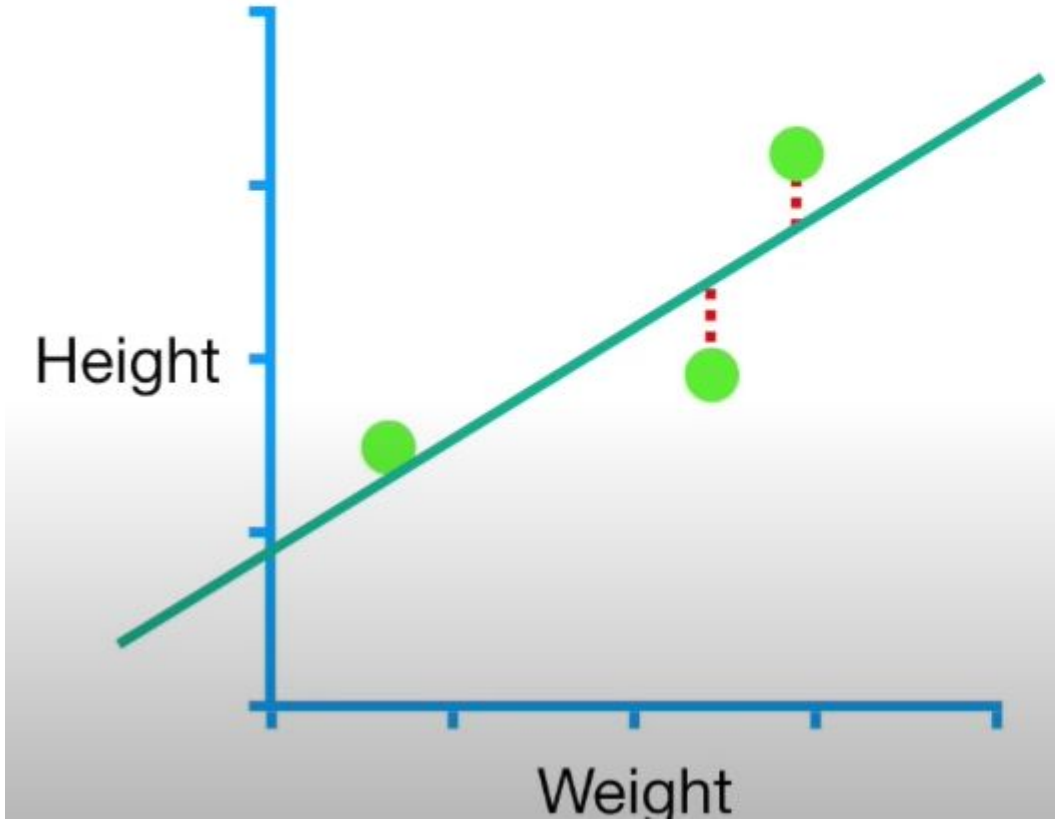
So if this **slope** = **0.009**. **Step Size** = **0.009** × **0.1** = 0.0009

…and get **0.0009**, which is smaller than **0.001**, so **Gradient Descent** would stop.

That said, **Gradient Descent** also includes a limit on the number of steps it will take before giving up.

In practice, the
Maximum Number of Steps = **1,000**
or greater.

Predicted Height = intercept + slope × **Weight**



…let's talk abut how to estimate the **Intercept** *and* the **Slope**.

$$\frac{d}{d\,slope}\,(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5}))^2 = 2(\mathbf{1.4} - (intercept + slope \times \mathbf{0.5})) \times \mathbf{-0.5}$$

$$\frac{d}{d\,slope}\,\mathbf{1.4} - (intercept + slope \times \mathbf{0.5})$$

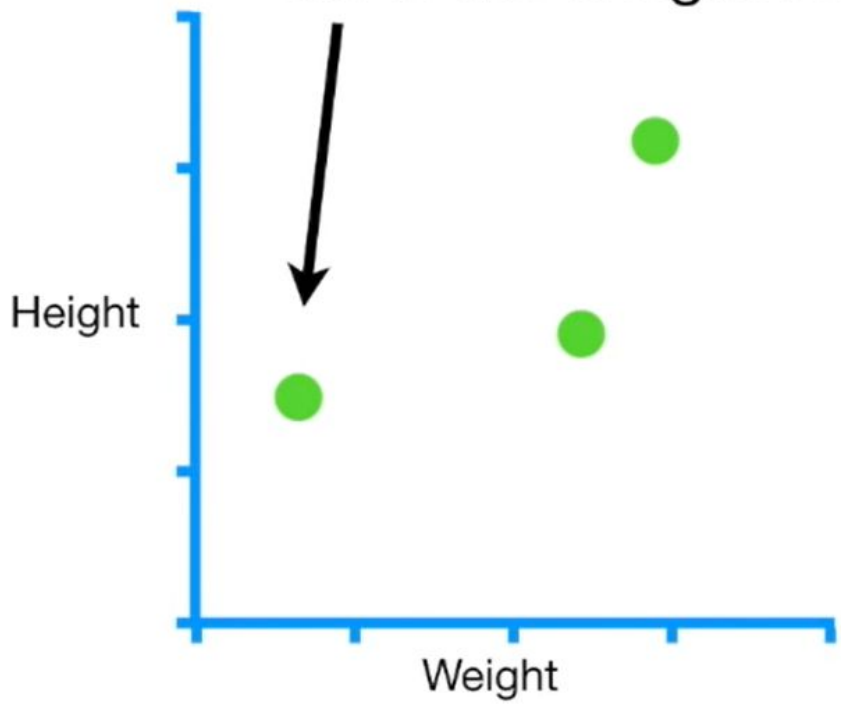
$$\frac{d}{d\,slope}\,\mathbf{1.4} + (-1)intercept - slope \times \mathbf{0.5}$$

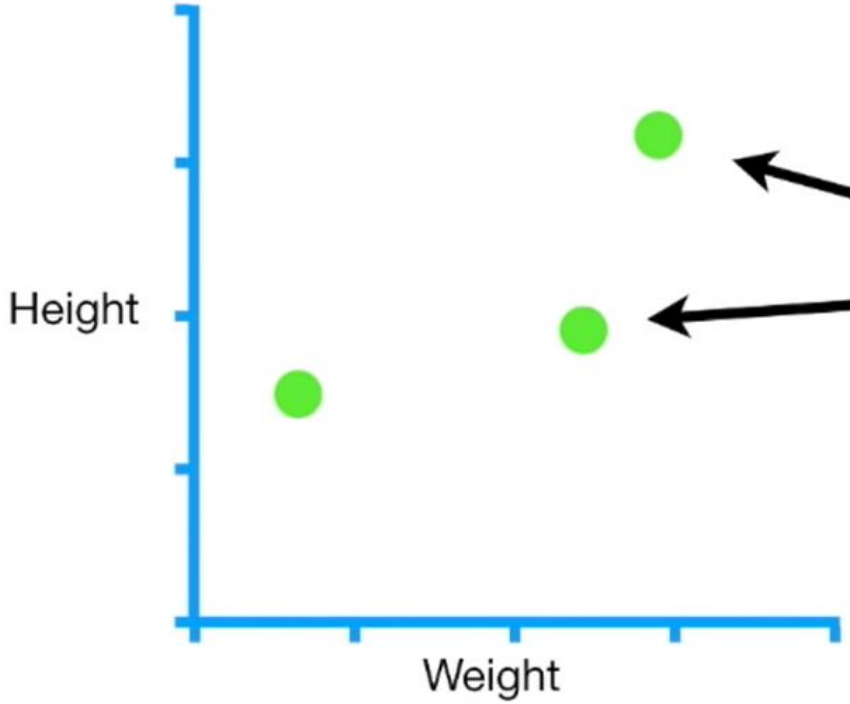
Since we are taking the derivative with respect to the **Slope**, we treat the **Intercept** like a constant, and the derivative of a constant is **0**.

$$\frac{d}{d\,slope}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))^2 = 2(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5})) \times \mathbf{-0.5}$$

$$= -2 \times \mathbf{0.5}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$

NOTE: I left the **0.5** in bold instead of multiplying it by 2 to remind us that **0.5** is the weight for the first sample.





Again, **2.3** and **2.9** are in bold to remind us that they are the weights of the second and third samples.

$$\frac{d}{d\,slope}\text{Sum of squared residuals} = -2 \times \mathbf{0.5}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$

$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$$

$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$

$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

$$-2(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))$$

We will use this **Gradient** to **descend** to lowest point in the **Loss Function**, which, in this case, is the Sum of the Squared Residuals…

…thus, this is why this algorithm is called **Gradient Descent**!

$$\frac{d}{d\ slope}$$ Sum of squared residuals =

$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (\text{intercept} + \text{slope} \times \mathbf{0.5}))$$

$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$

$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$$

Just like before, we will start by picking a random number for the **Intercept**. In this case we'll set the **Intercept** = **0**…

…and we'll pick a random number for the **Slope**. In this case we'll set the **Slope** = **1**.

Now let's plug in **0** for the **Intercept** and **1** for the **Slope**…

$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

$$-2(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9})) = \mathbf{-1.6}$$

Step Size$_{\text{Intercept}}$ = **Slope** × **Learning Rate**

…now we plug the **Slopes** into the **Step Size** formulas…

$$\frac{d}{d\,slope}$$ Sum of squared residuals =

$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5}))$$

$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9}))^2$$

$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3}))^2 = \mathbf{-0.8}$$

Step Size$_{\text{Slope}}$ = **Slope** × **Learning Rate**

$$\frac{d}{d\ intercept}\ \text{Sum of squared residuals} =$$

$$-2(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5})$$

$$+ -2(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3}))$$

$$+ -2(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9})) = \mathbf{-1.6}$$

Step Size$_{\text{Intercept}}$ = -1.6 x 0.01 = **-0.016**

New Intercept = 0 - (-0.016)

…and the
Step Sizes…

$$\frac{d}{d\ slope}\ \text{Sum of squared residuals} =$$

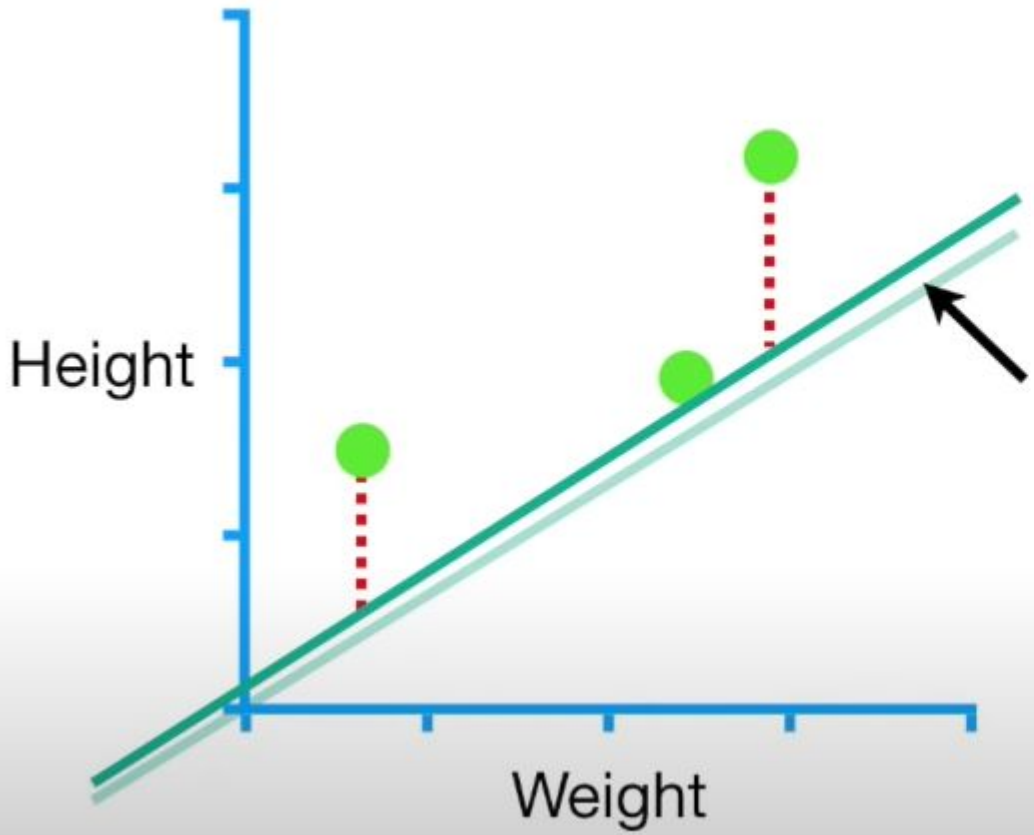
$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5}))$$

$$+ -2 \times \mathbf{2.9}(\mathbf{3.2} - (0 + 1 \times \mathbf{2.9}))^2$$

$$+ -2 \times \mathbf{2.3}(\mathbf{1.9} - (0 + 1 \times \mathbf{2.3}))^2 = \mathbf{-0.8}$$

Step Size$_{\text{Slope}}$ = -0.8 x 0.01 = **-0.008**

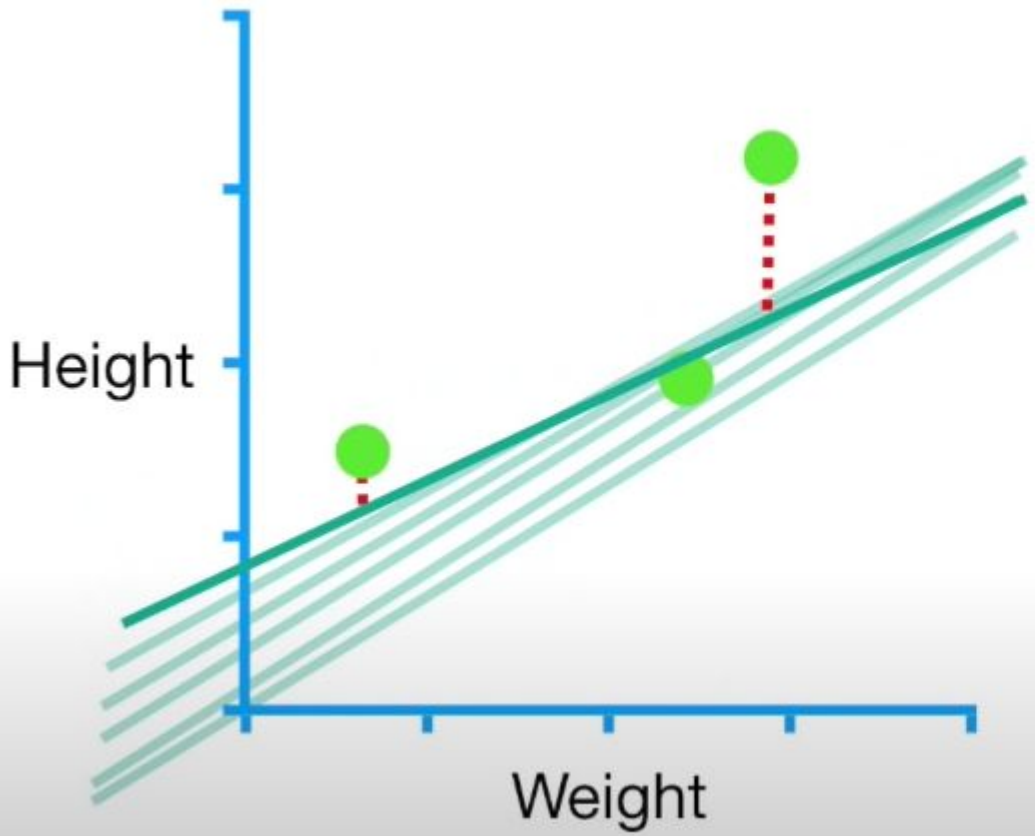
New Slope = 1 - (-0.008)



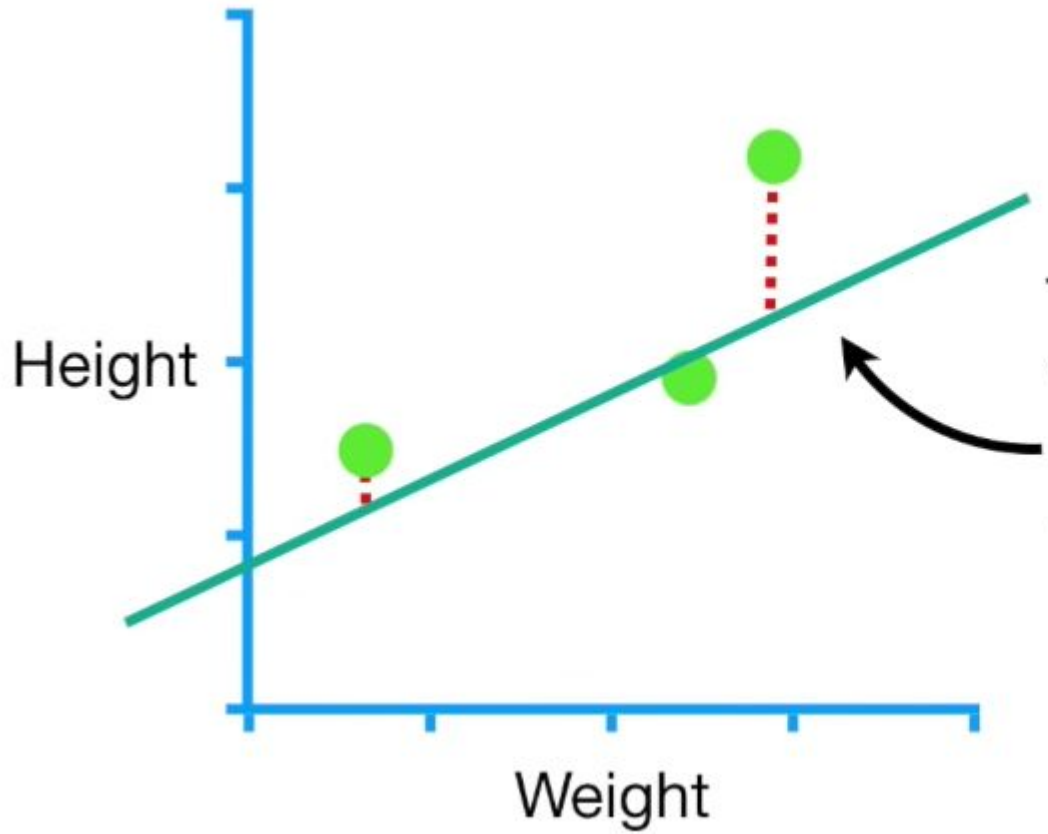
New Intercept = 0 - (-0.016) = 0.016

…and this is the new line (with **Slope = 1.008** and **Intercept = 0.016**) after the first step.

New Slope = 1 - (-0.008) = 1.008



Now we just repeat what we did until all of the **Steps Sizes** are very small or we reach the **Maximum Number of Steps**.



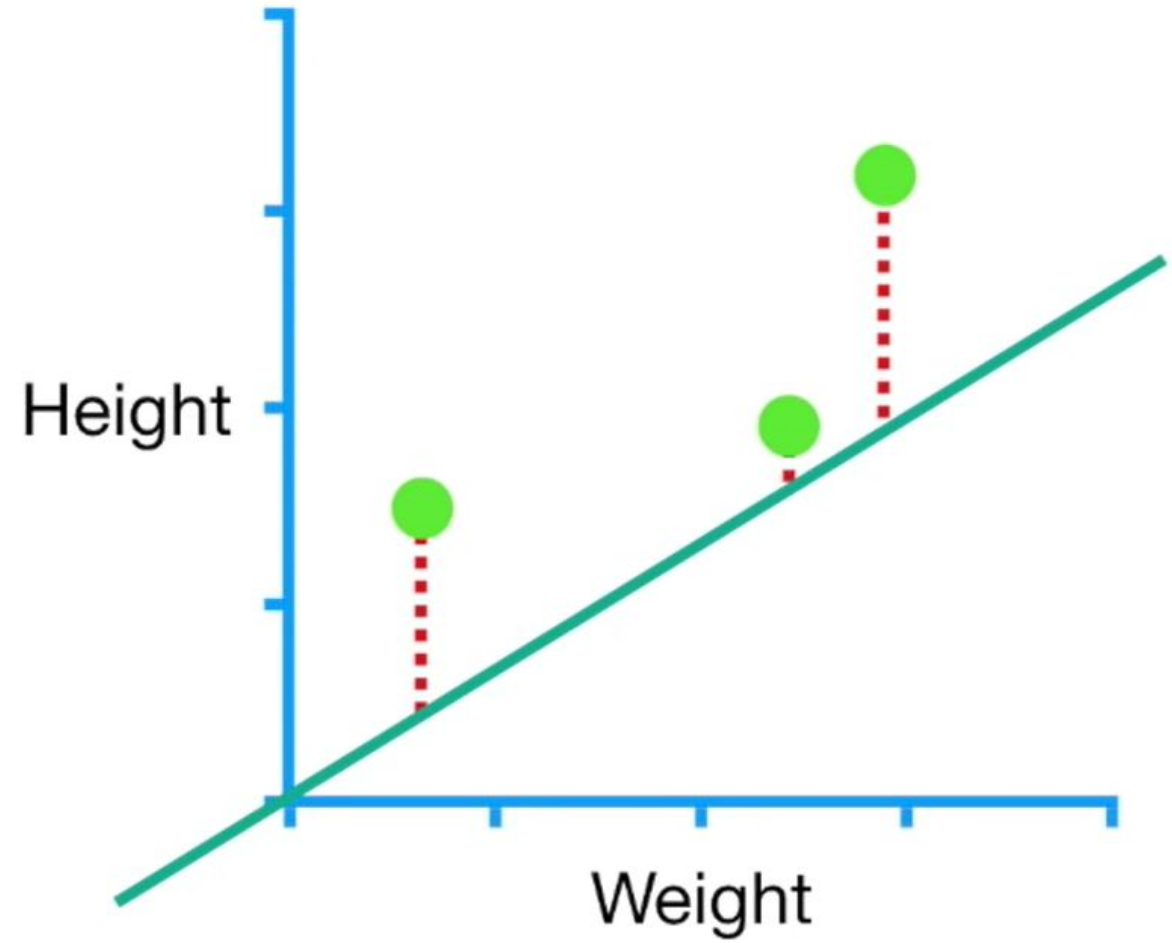
This is the best fitting line, with **Intercept** = **0.95** and **Slope** = **0.64**, the same values we get from **Least Squares**.

Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

+ (**1.9** - (intercept + 0.64 × **2.3**))^2

+ (**3.2** - (intercept + 0.64 × **2.9**))^2

NOTE: The Sum of the Squared Residuals is just one type of **Loss Function**.



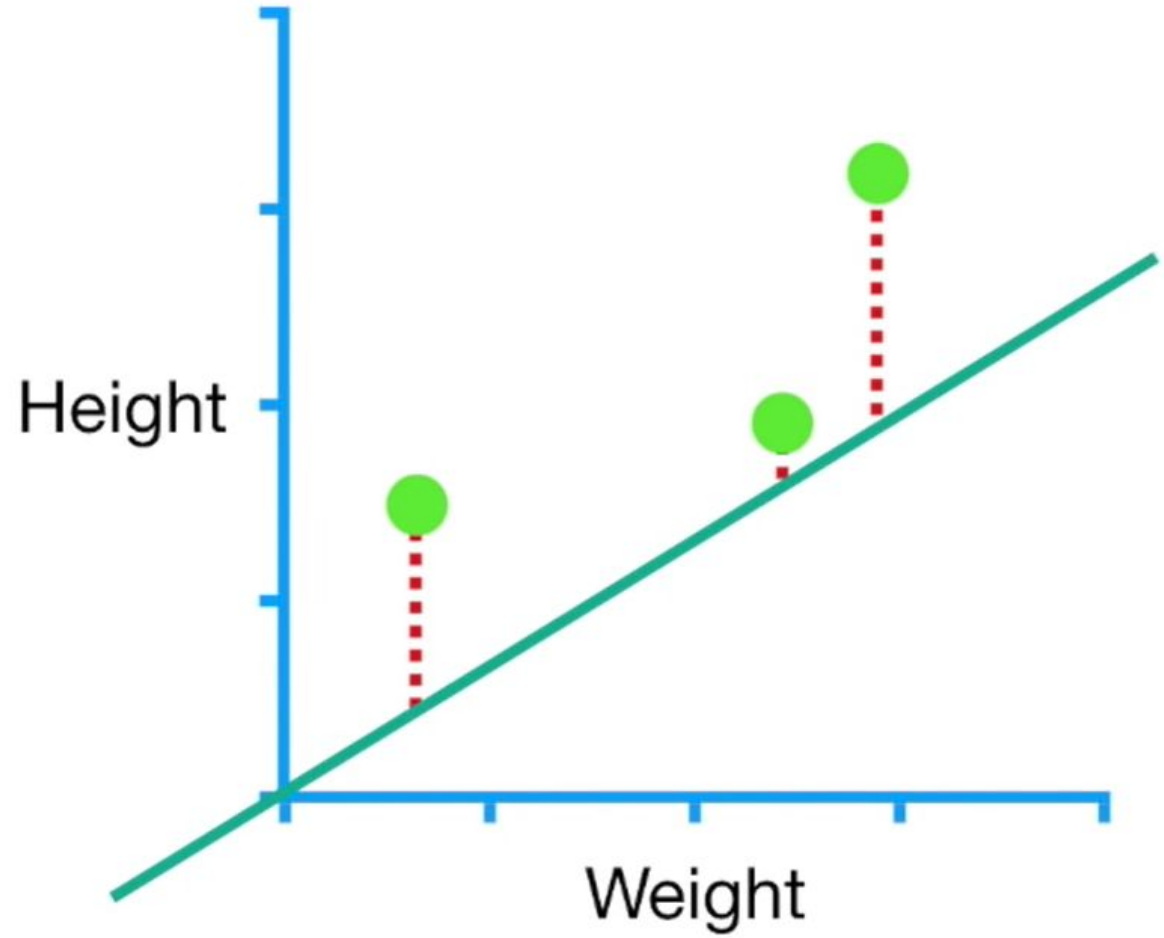
Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

+ (**1.9** - (intercept + 0.64 × **2.3**))^2

+ (**3.2** - (intercept + 0.64 × **2.9**))^2

However, there are tons of other **Loss Functions** that work with other types of data.

Regardless of which **Loss Function** you use, **Gradient Descent** works the same way.



Step 1: Take the derivative of the **Loss Function** for each parameter in it.
In fancy Machine Learning Lingo, take the **Gradient** of the **Loss Function**.

Step 2: Pick random values for the parameters.

Step 3: Plug the parameter values into the derivatives (ahem, the **Gradient**).

Step 4: Calculate the Step Sizes: **Step Size** = **Slope** × **Learning Rate**

Step 5: Calculate the New Parameters:

New Parameter = **Old Parameter** - **Step Size**

Now go back to **Step 3** and repeat until
Step Size is very small, or you reach
the **Maximum Number of Steps**.

Reference

https://www.youtube.com/watch?v=sDv4f4s2SB8