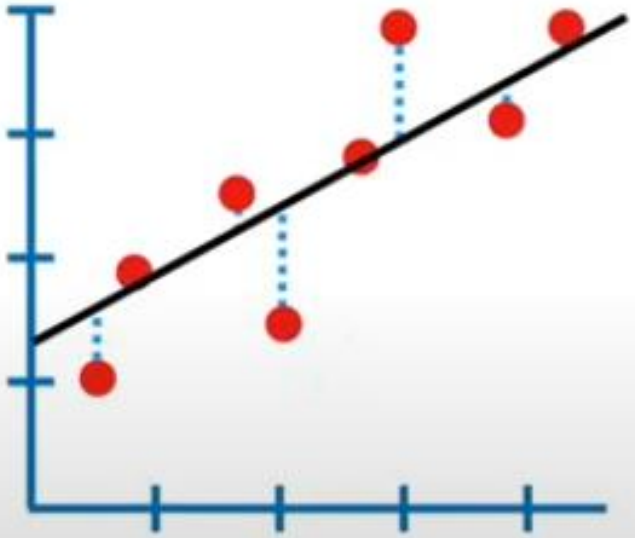
Gradient Descent Step-by-Step!!!

NOTE: This StatQuest assumes you already understand the basics of **Least Squares** and **Linear Regression**, so if you're not already down with that, check out the Quest…

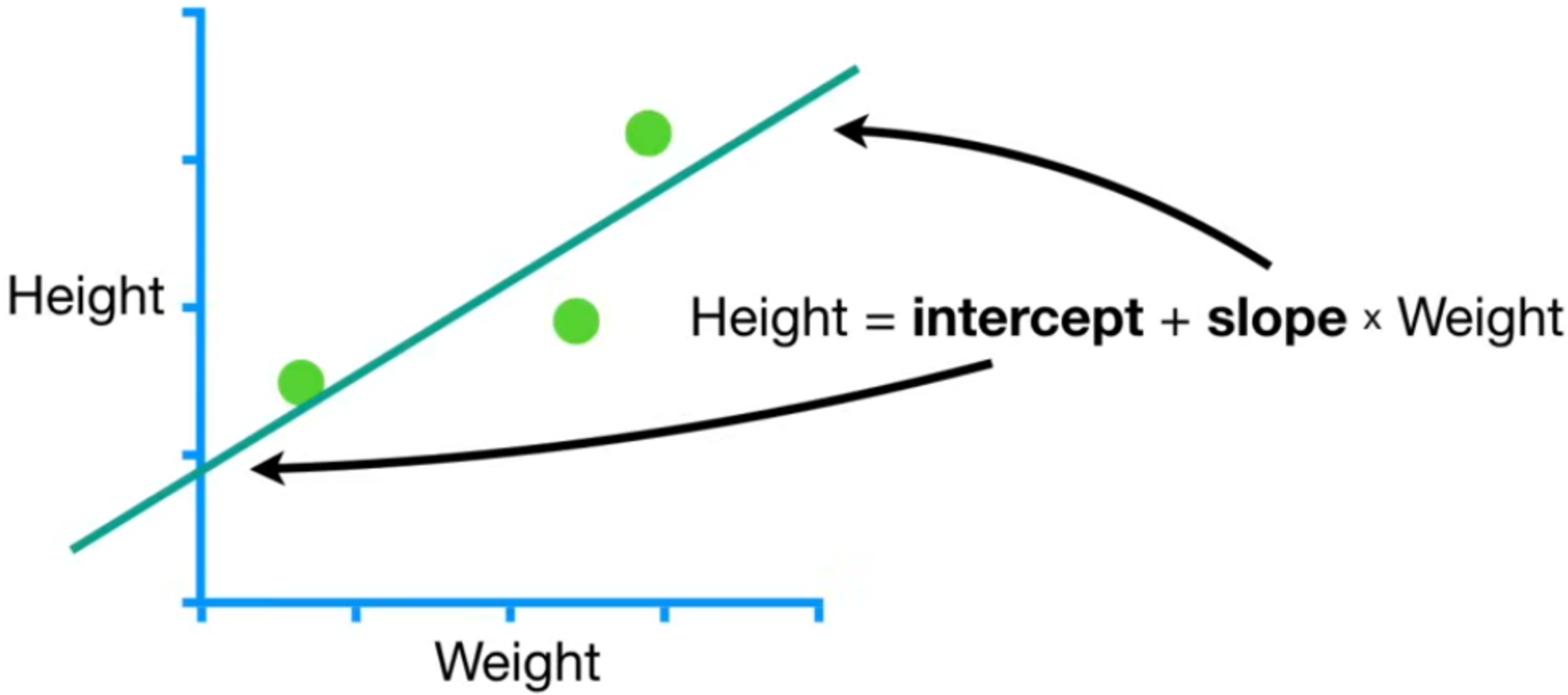
Fitting a Line to Data…



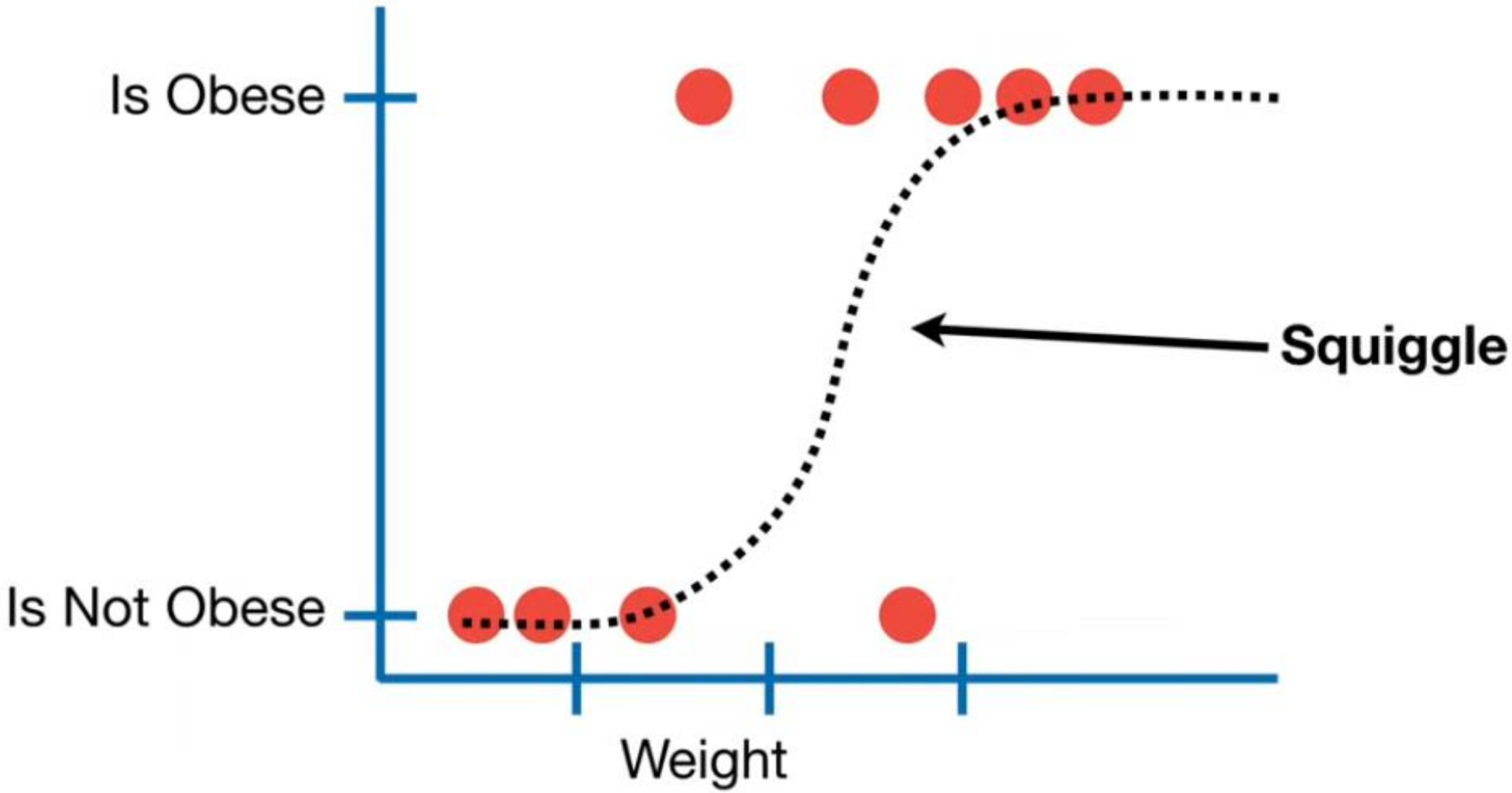
aka Linear Regression

In Statistics, Machine Learning and other Data Science fields, we optimize a lot of stuff.

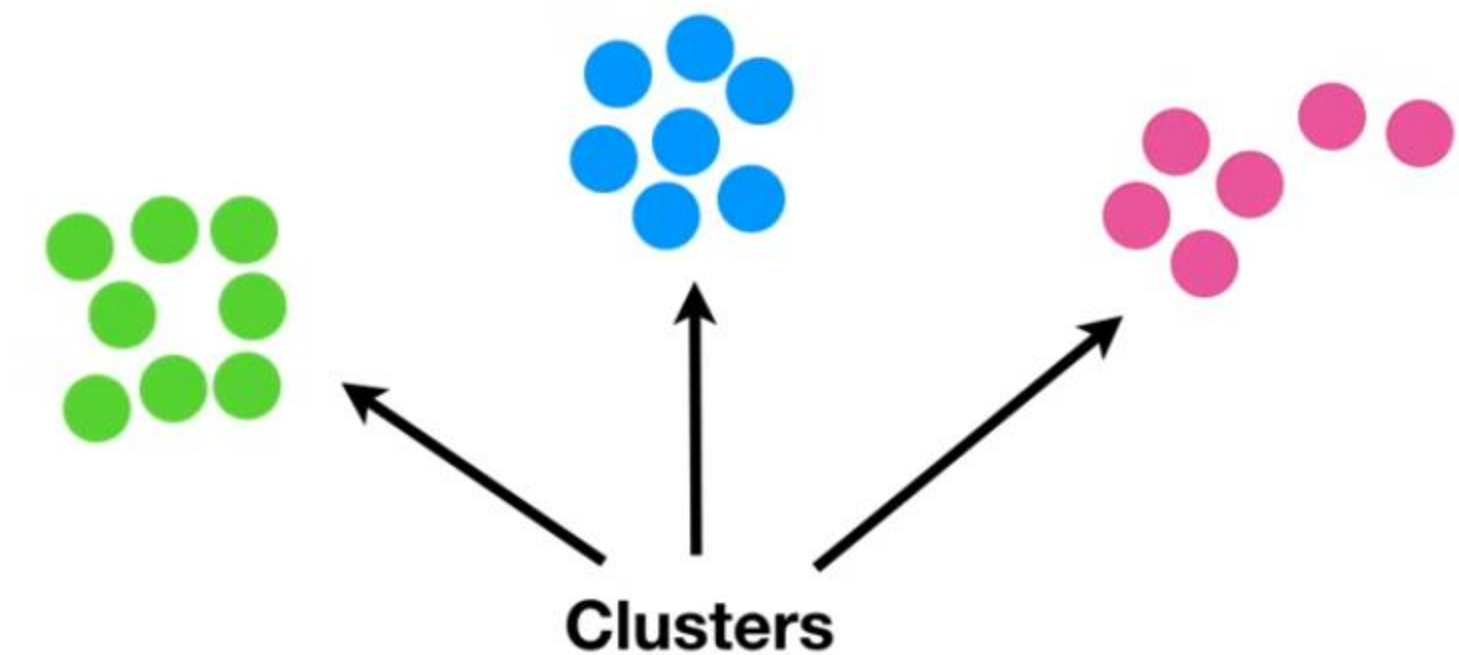
When we fit a line with **Linear Regression**, we optimize the **Intercept** and **Slope**.



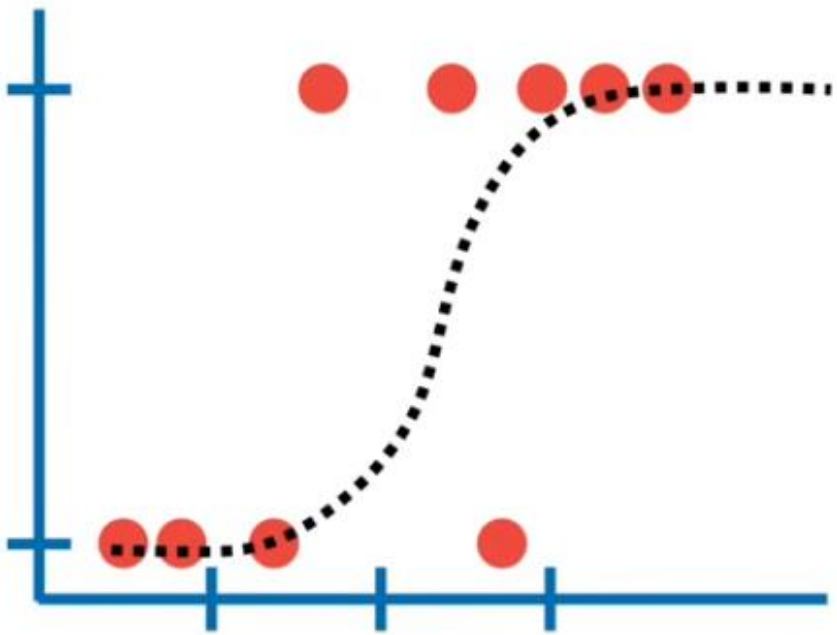
When we use **Logistic Regression**,
we optimize a squiggle.

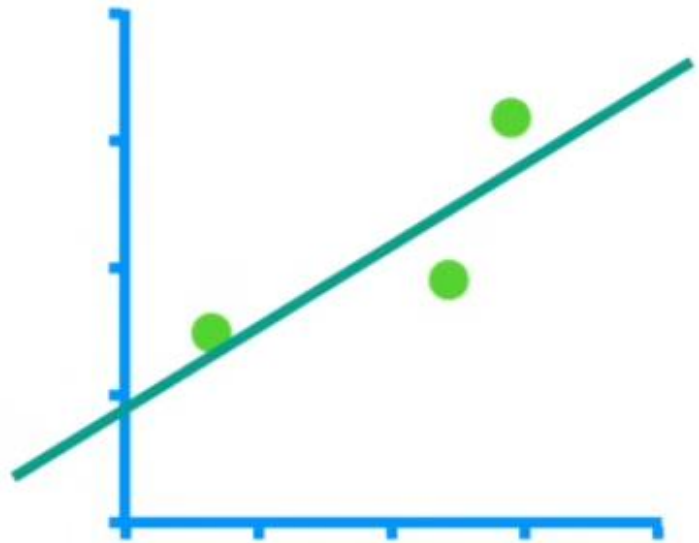


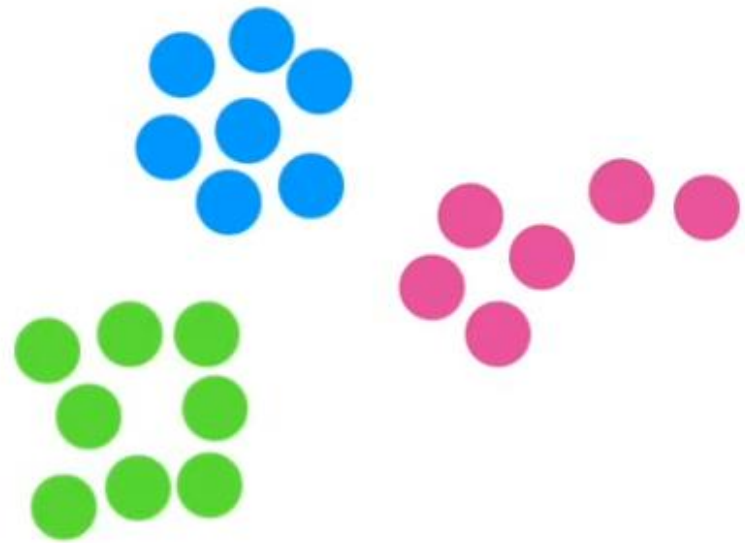
And when we use **t-SNE**, we optimize clusters.



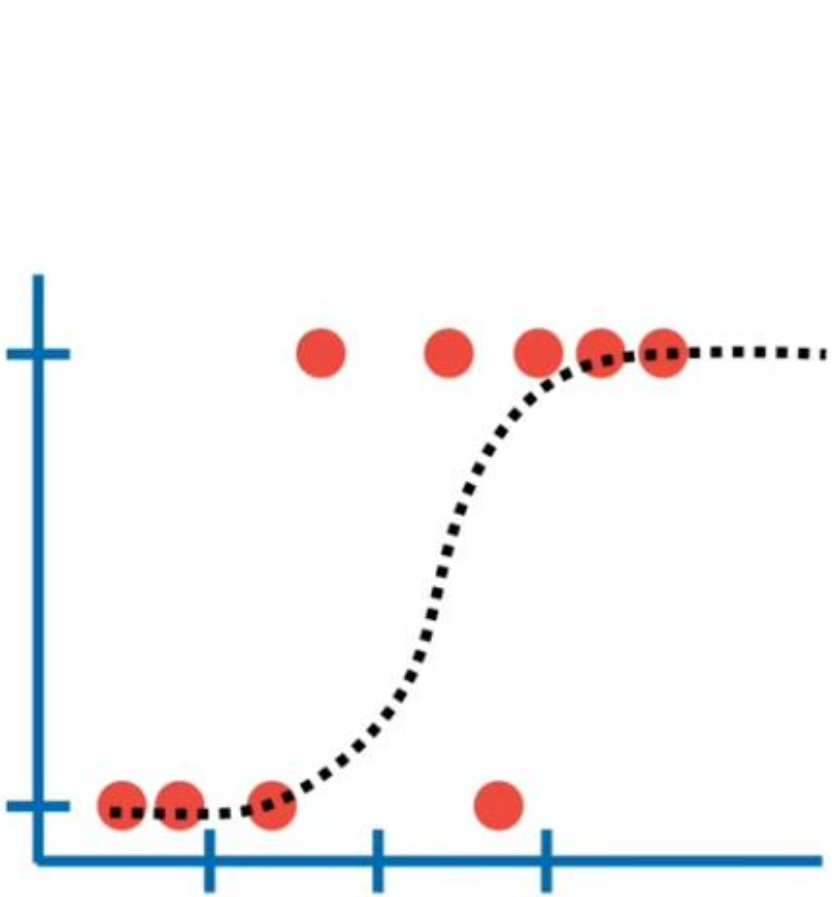
These are just a few examples of the stuff we optimize, there are tons more.

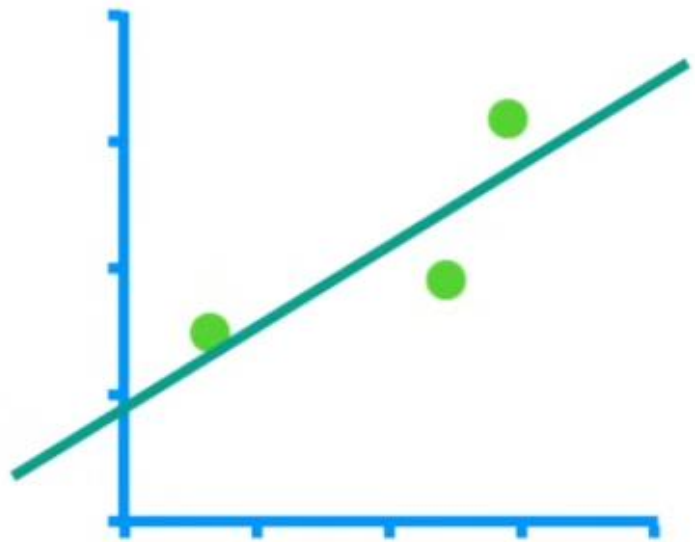


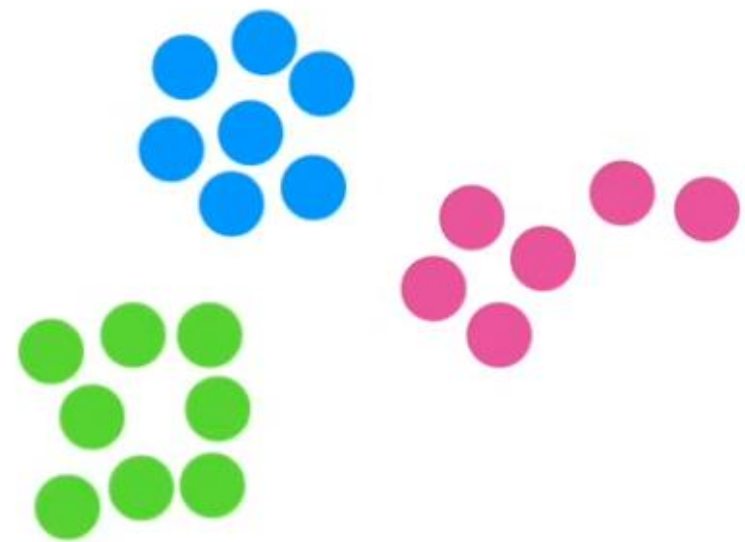




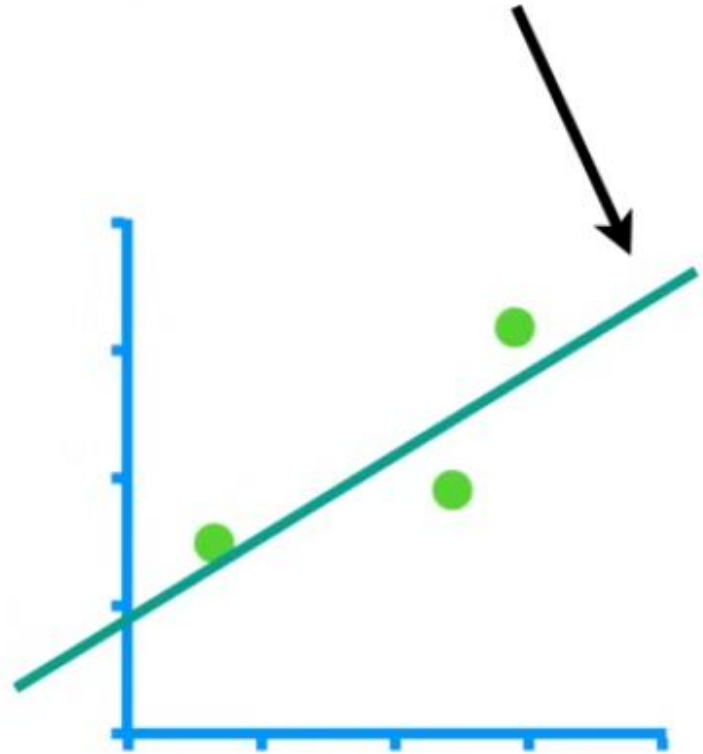
The cool thing is that **Gradient Descent** can optimize all these things and much more.



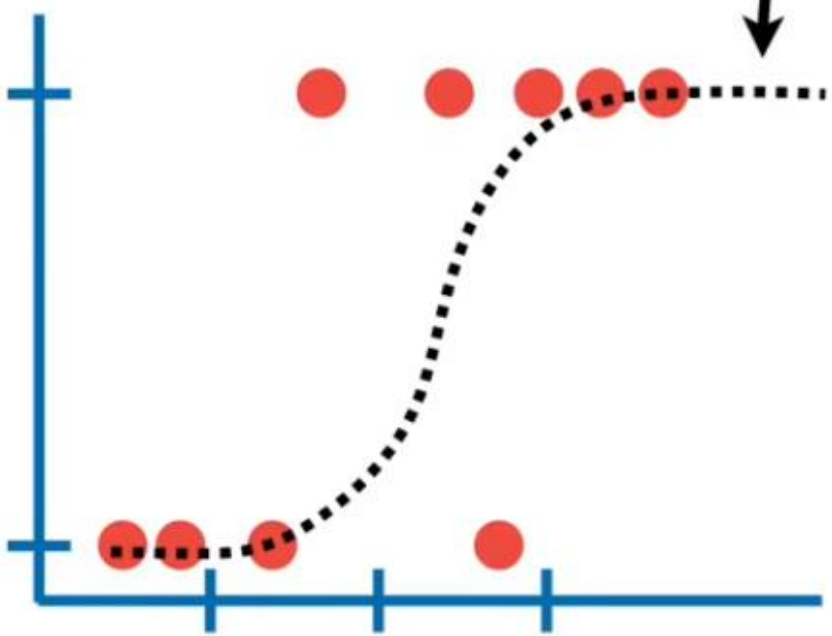




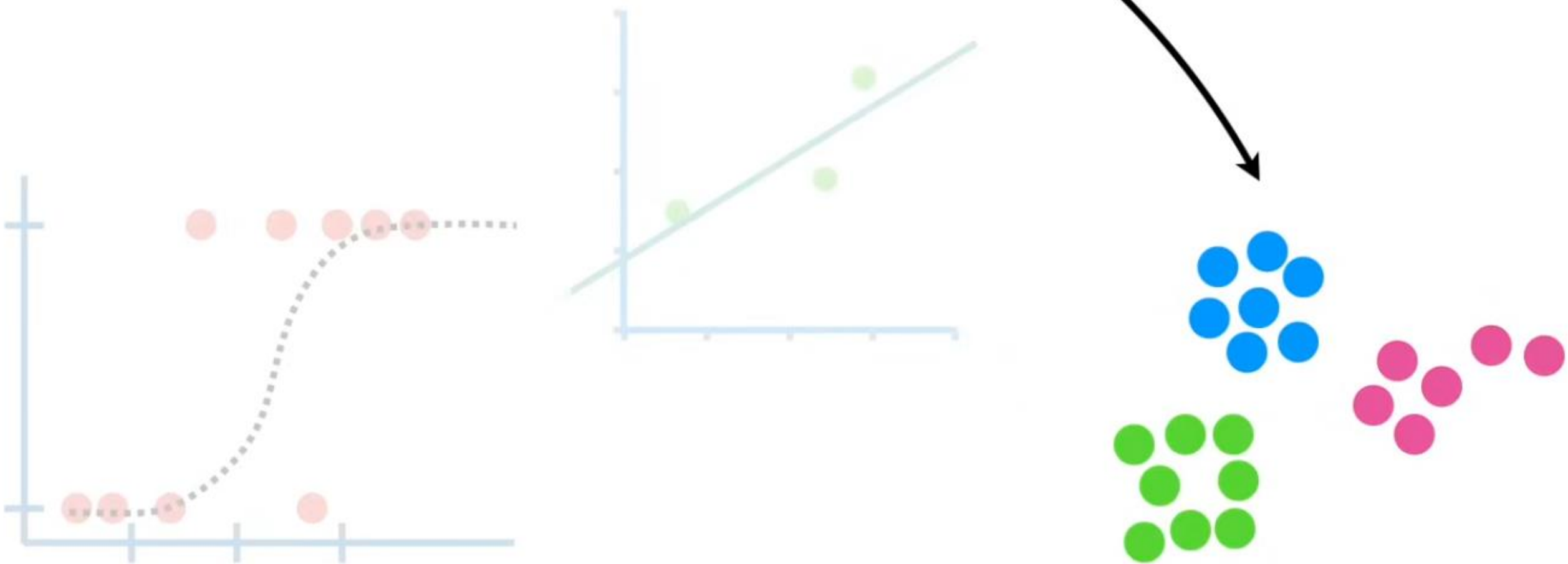
So if we learn how to optimize this line using **Gradient Descent**…



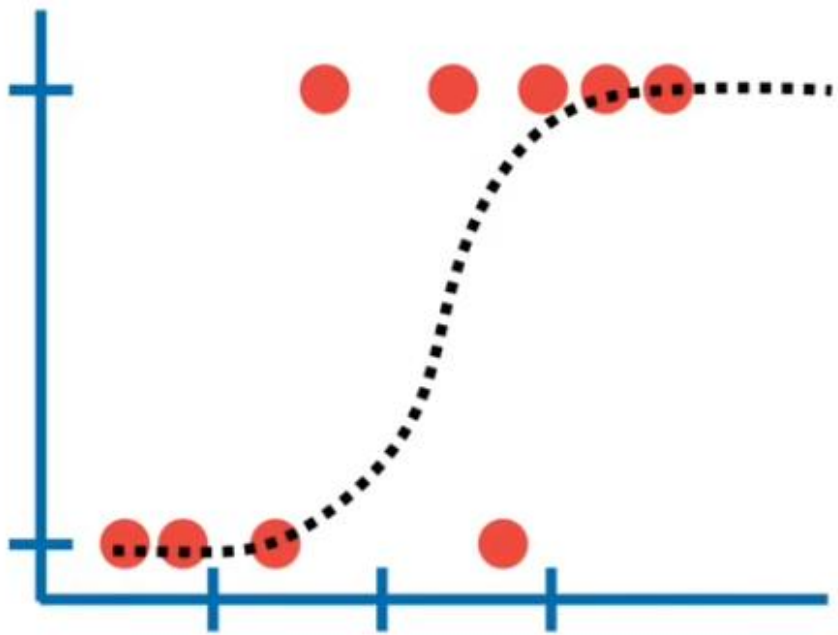
…then we will have learned the strategy that optimizes this squiggle…

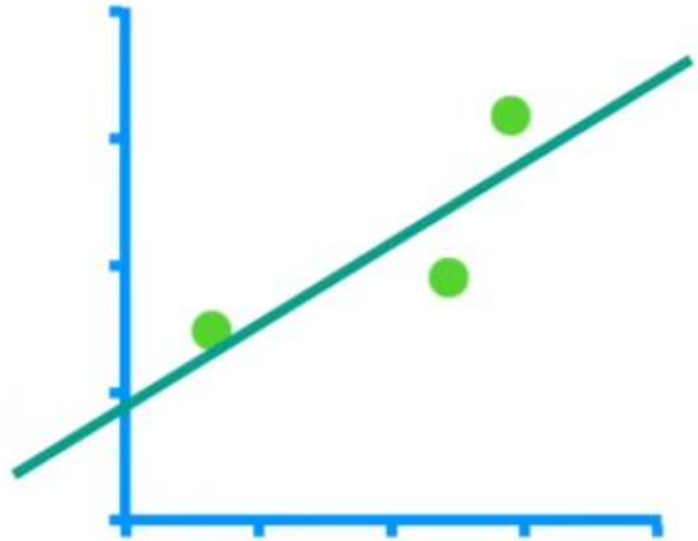


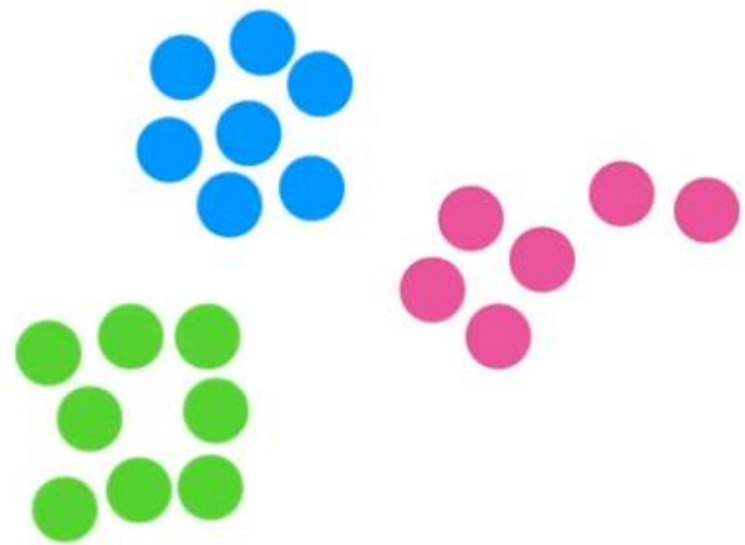
…and these clusters…

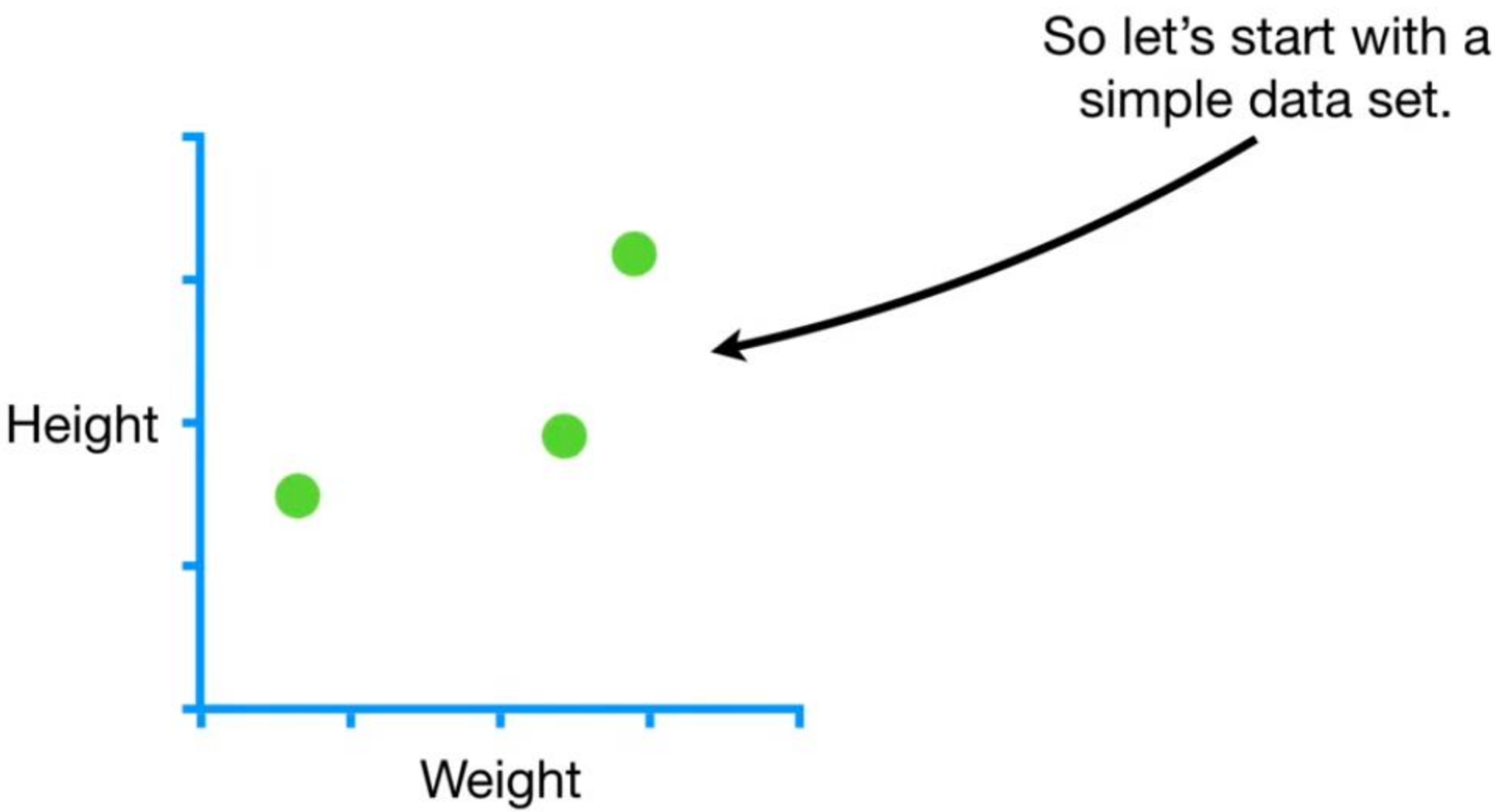


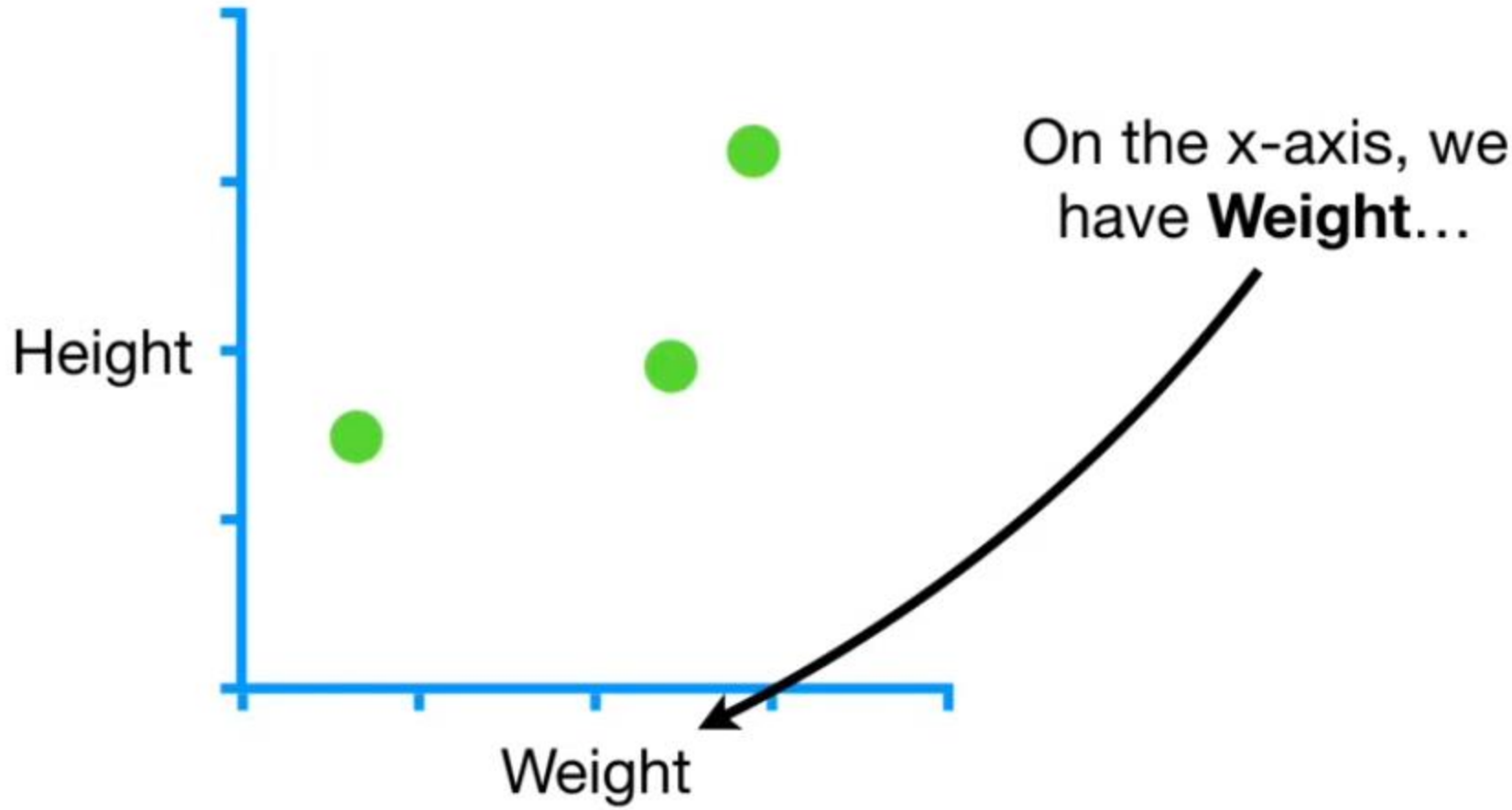
…and many more of the optimization problems we have in Statistics, Machine Learning and Data Science.

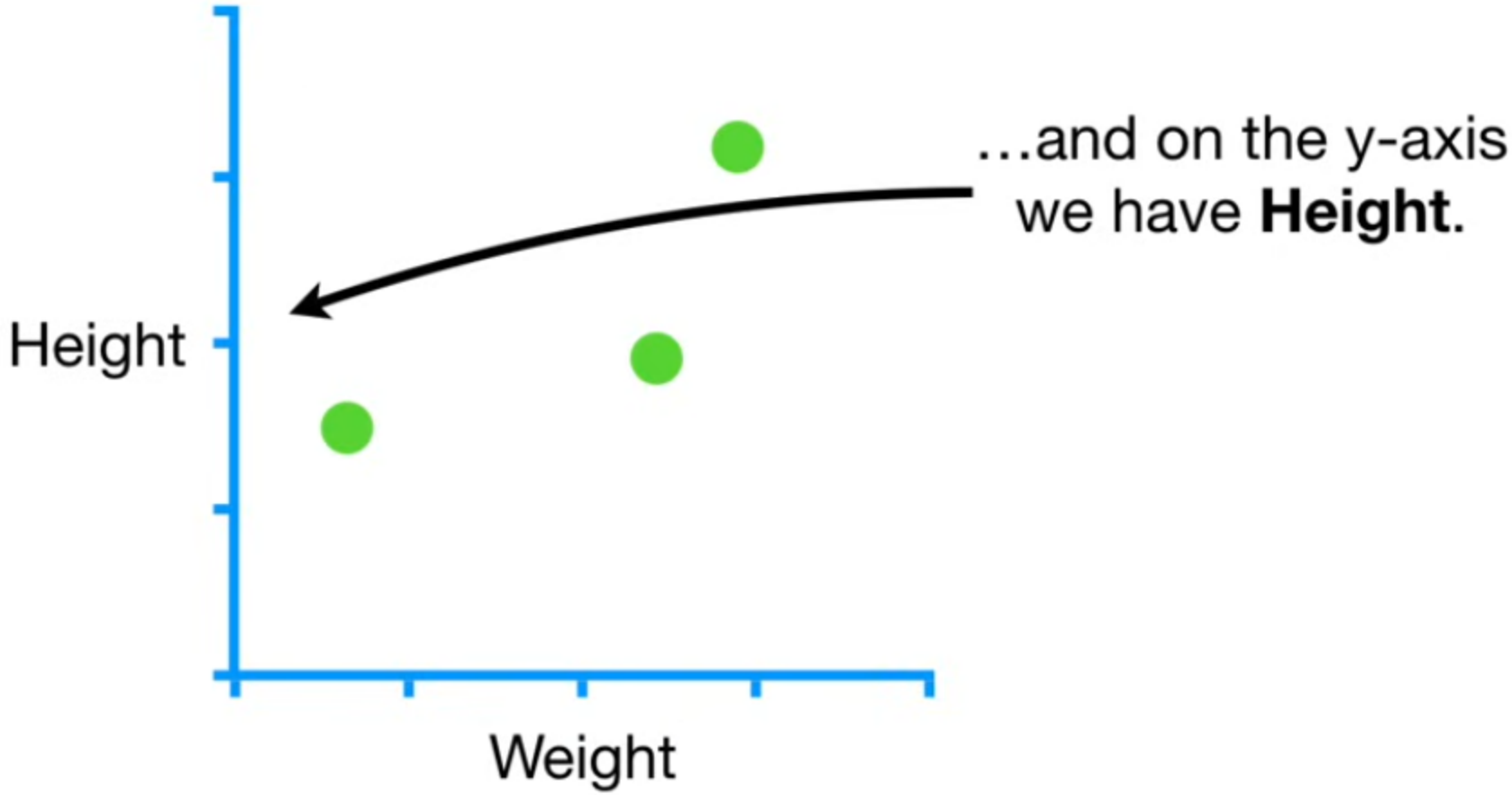


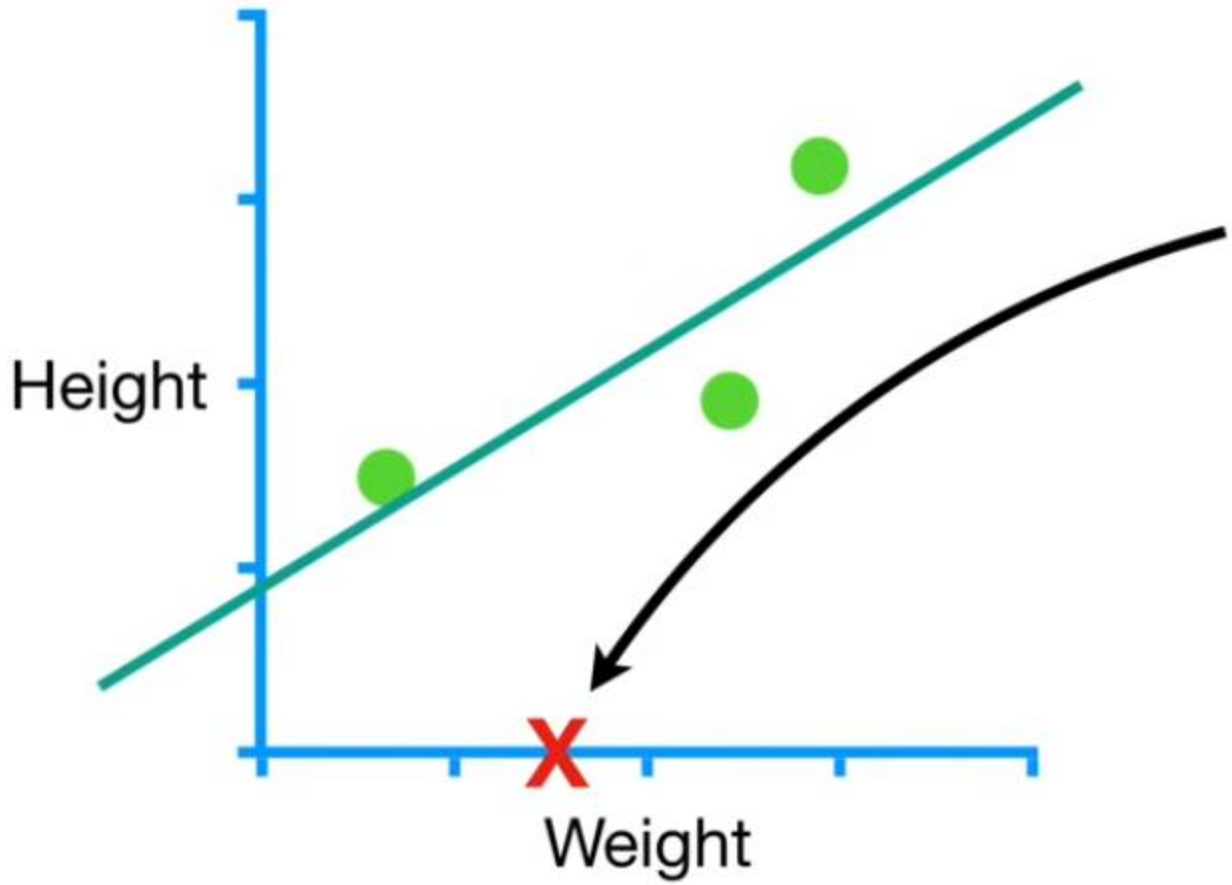




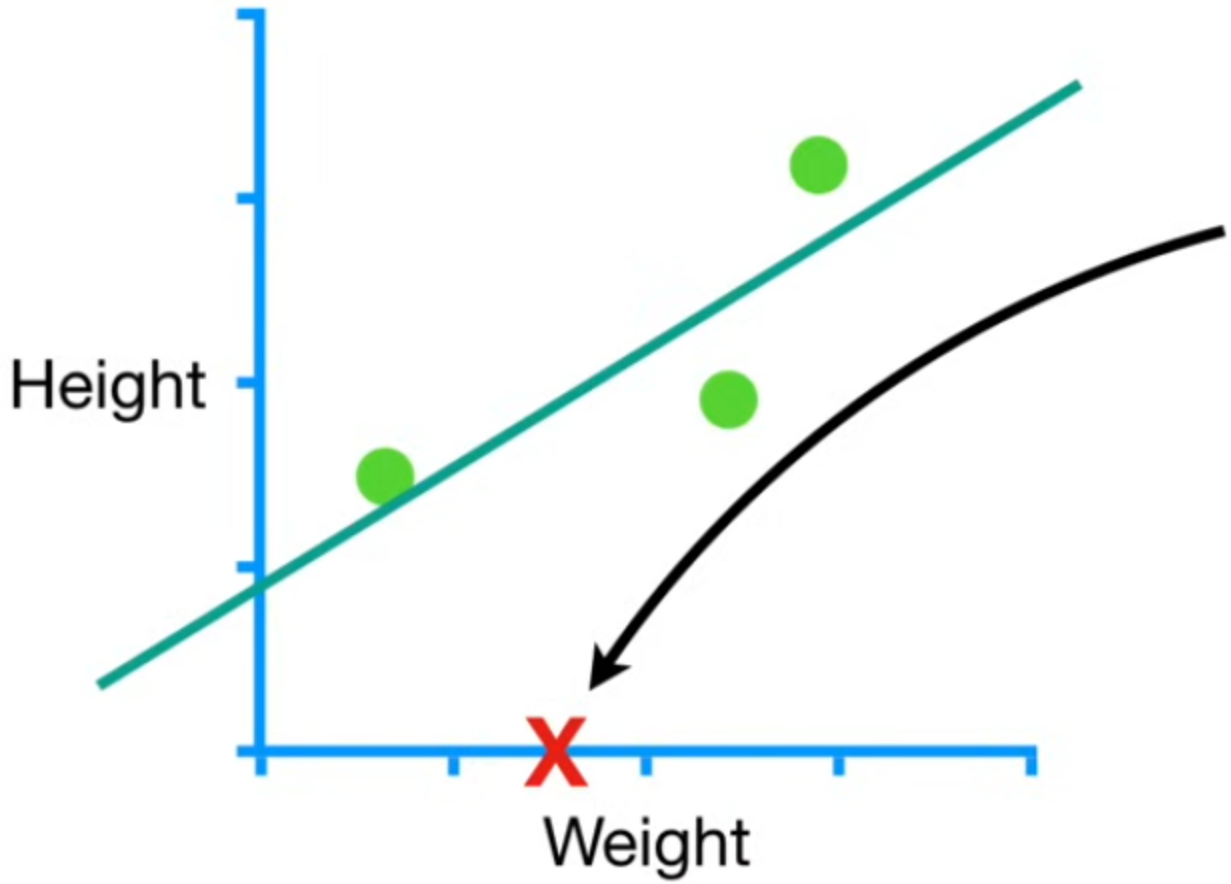

So let's start with a
simple data set.
Height
Weight


Height
Weight
On the x-axis, we have **Weight**…


Height
Weight
…and on the y-axis
we have **Height**.

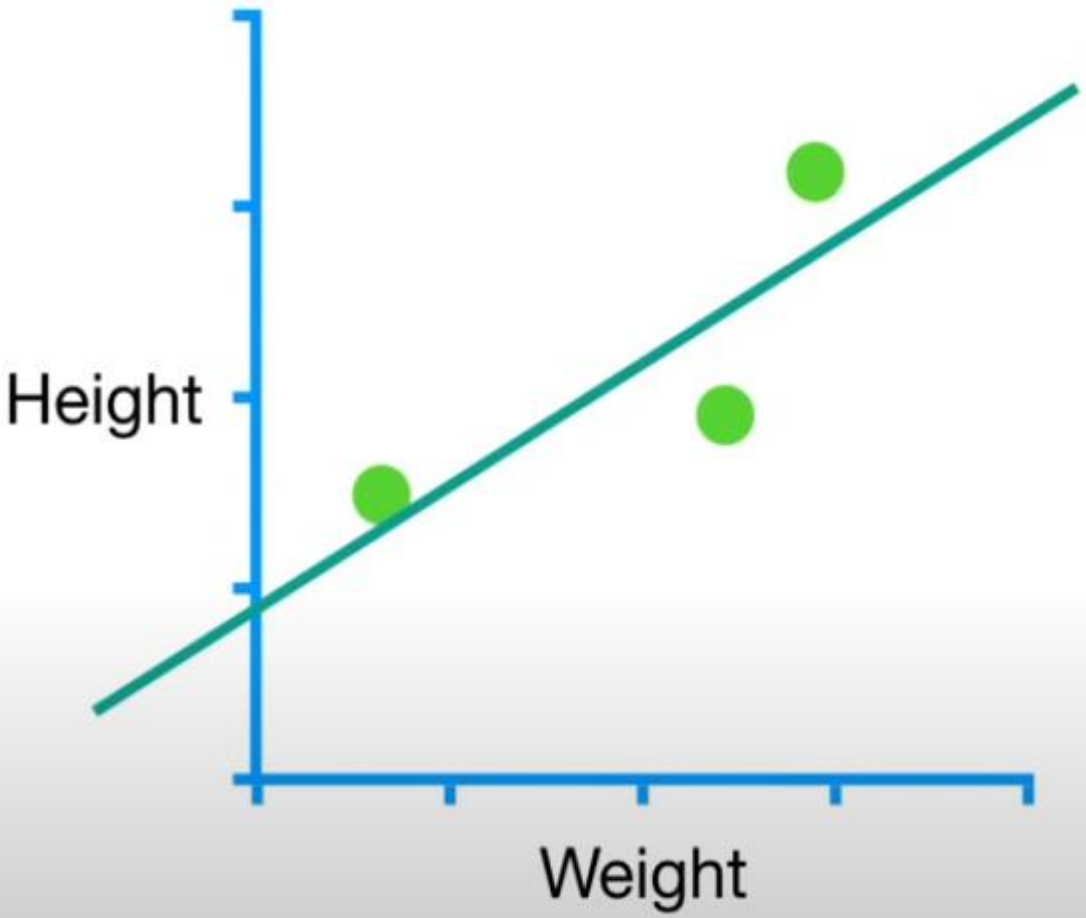


…and someone tells us
that they weigh **1.5**…



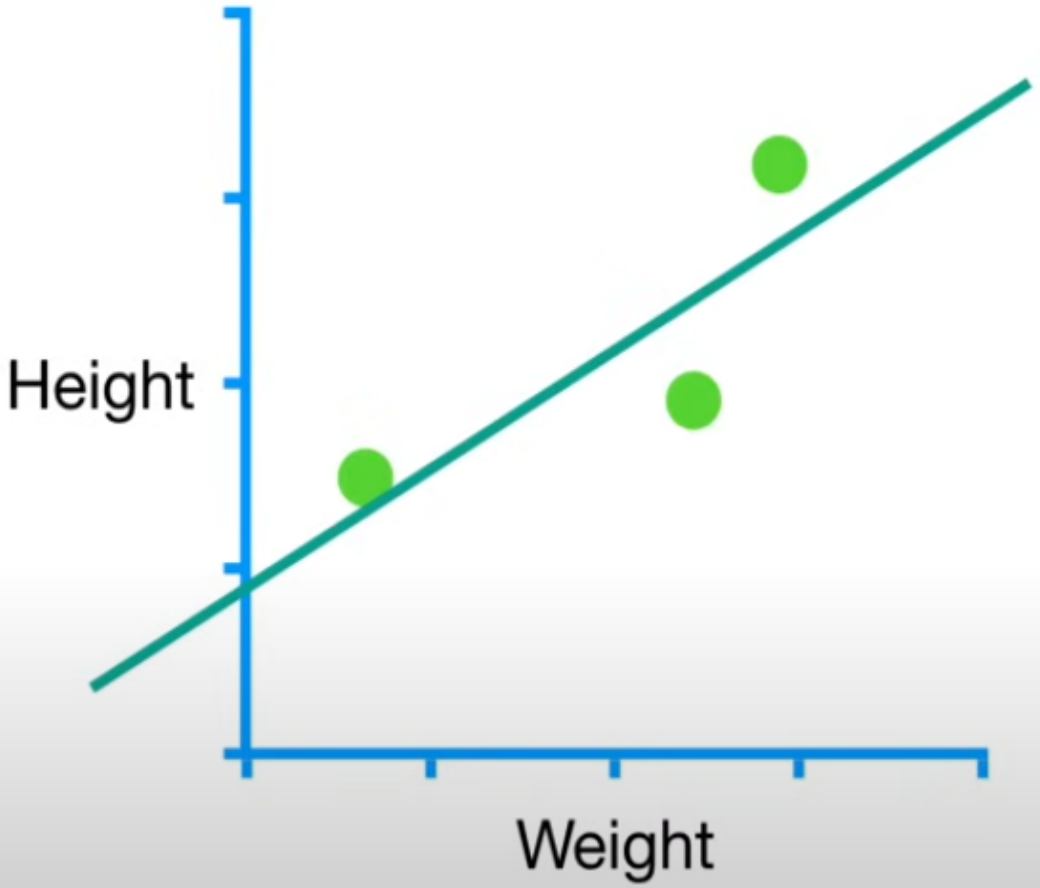
…and someone tells us
that they weigh **1.5**…

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



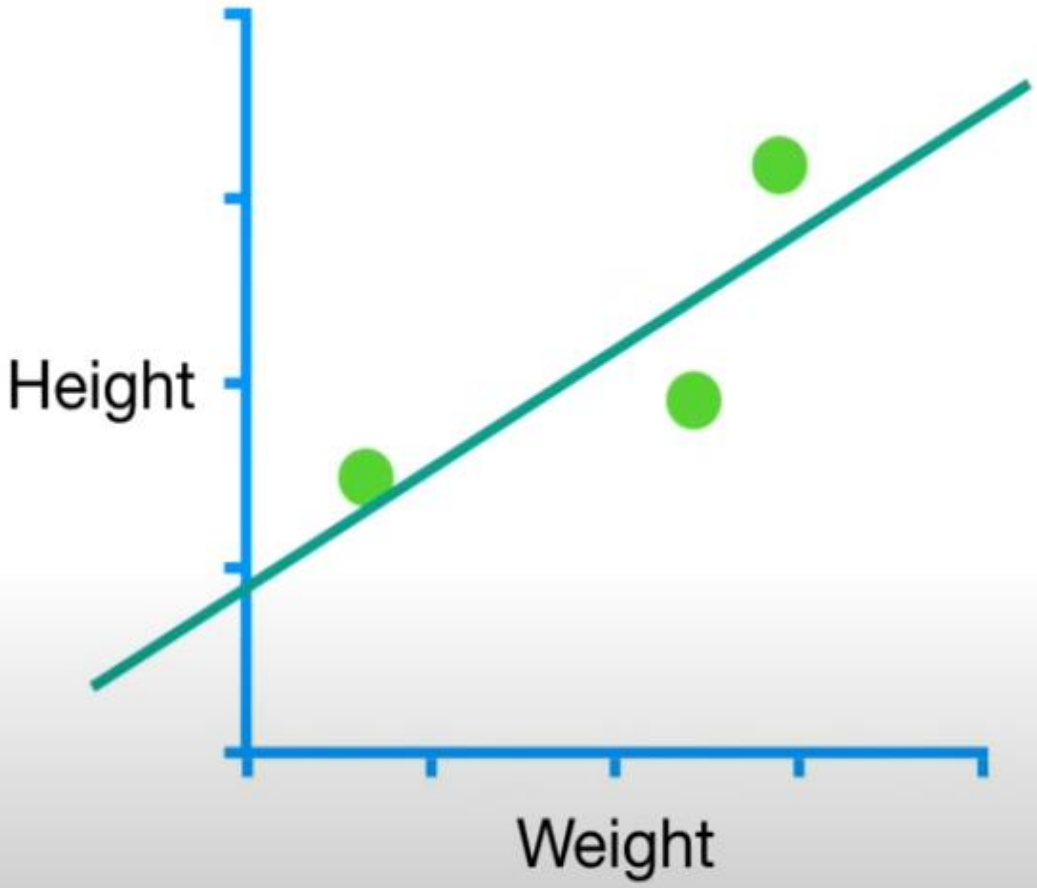
So let's learn how **Gradient Descent** can fit a line to data by finding the optimal values for the **Intercept** and the **Slope**.

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



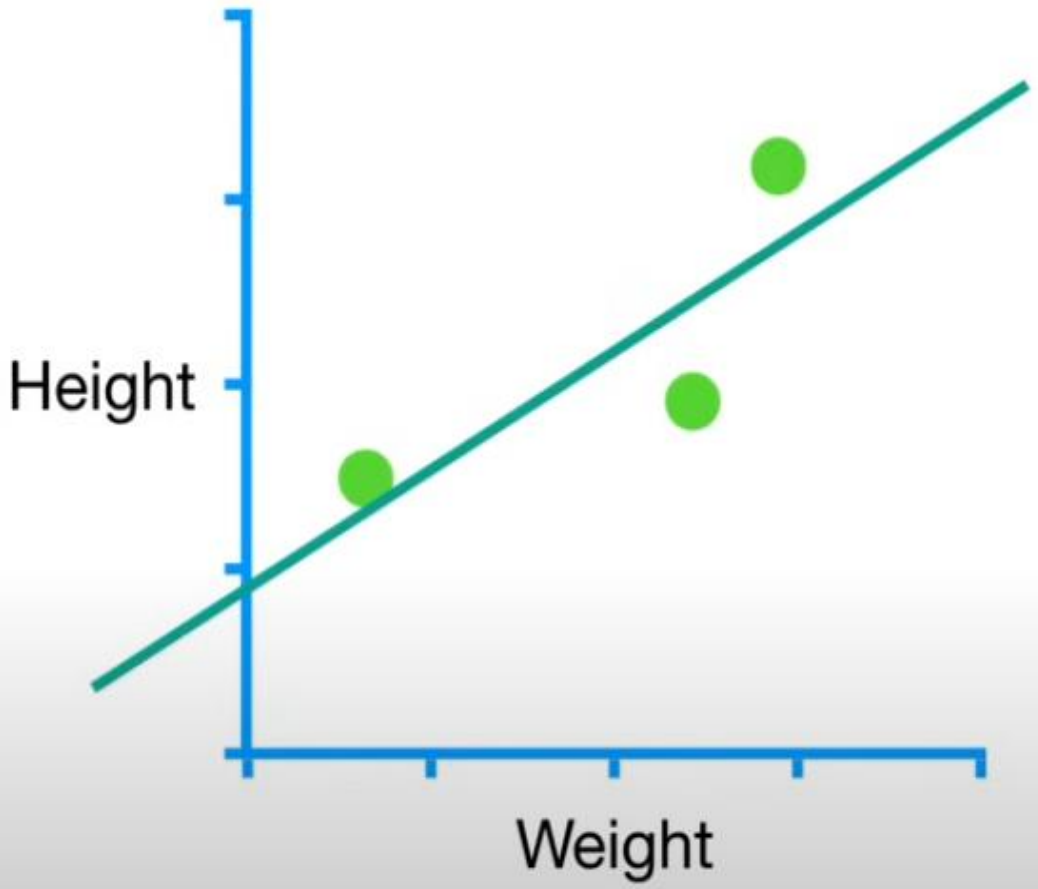
Actually, we'll start by using **Gradient Descent** to find the **Intercept**.

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



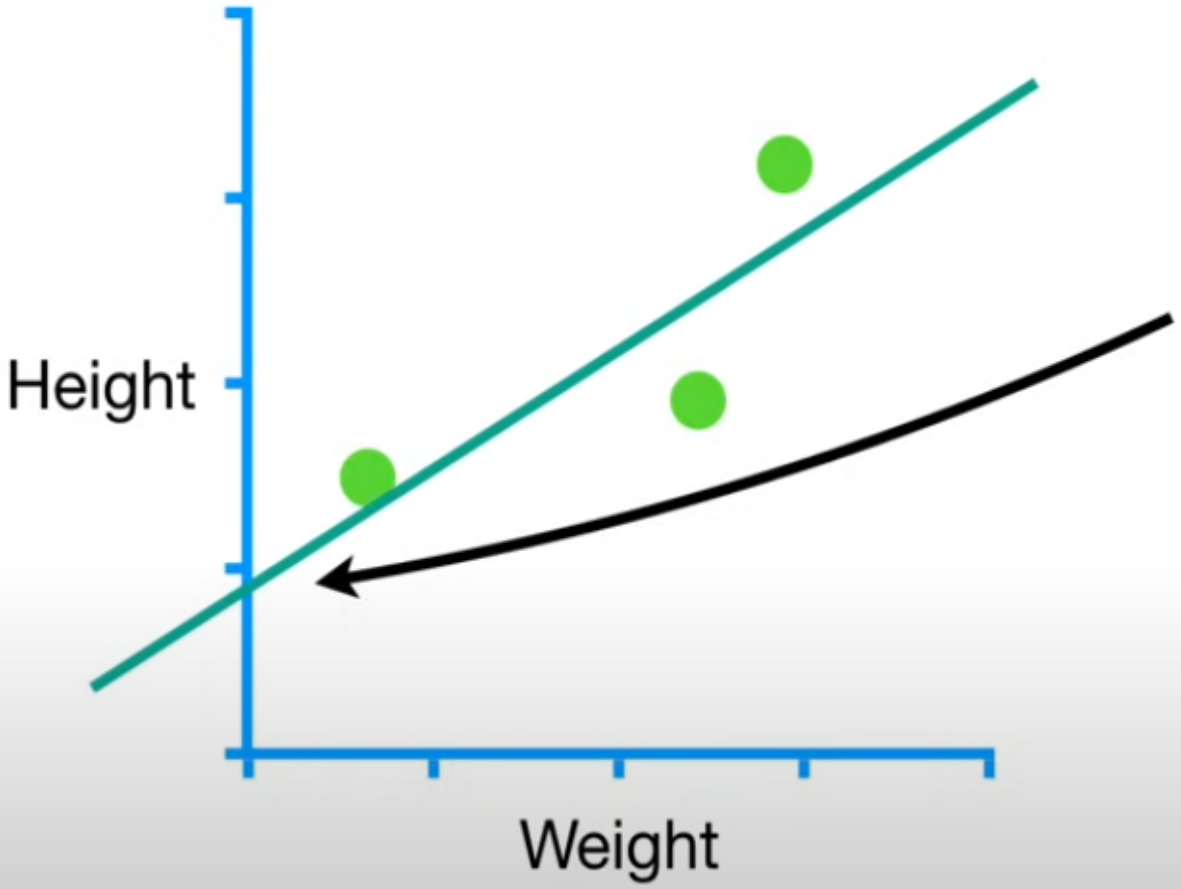
Then, once we understand how **Gradient Descent** works, we'll use it to solve for the **Intercept** *and* the **Slope**.

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



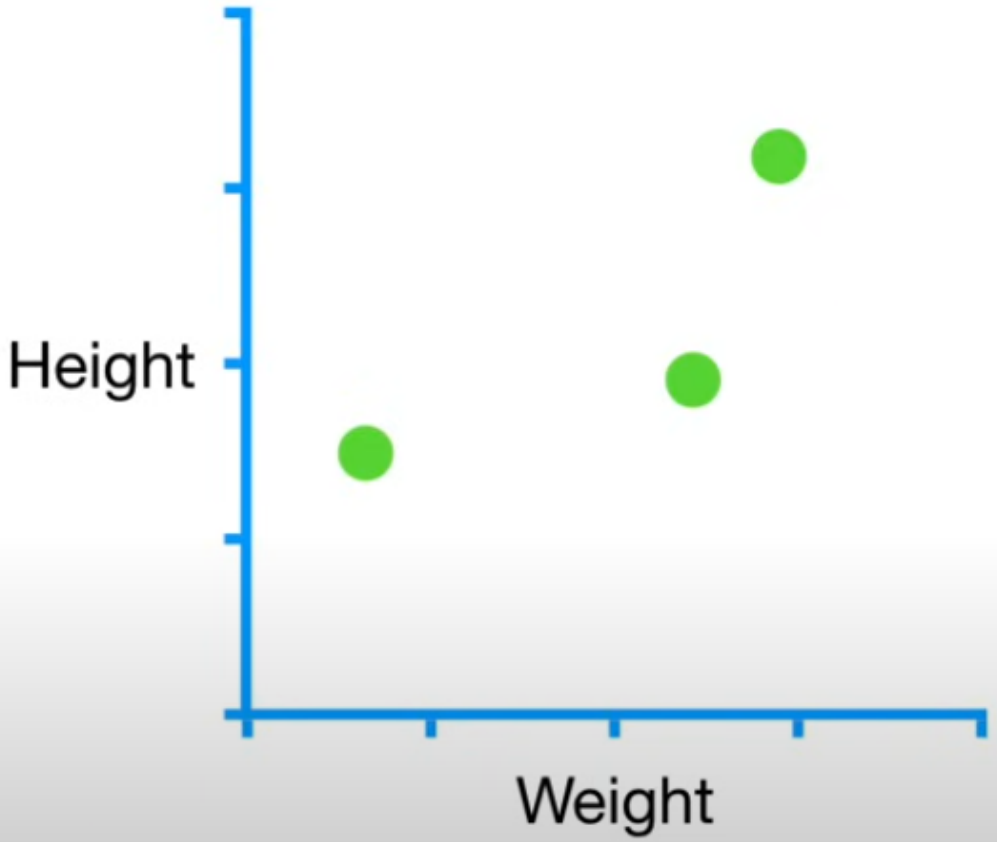
So for now, let's just plug in the **Least Squares** estimate for the **Slope**, **0.64**.

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



…and we'll use **Gradient Descent** to find the the optimal value for the Intercept.

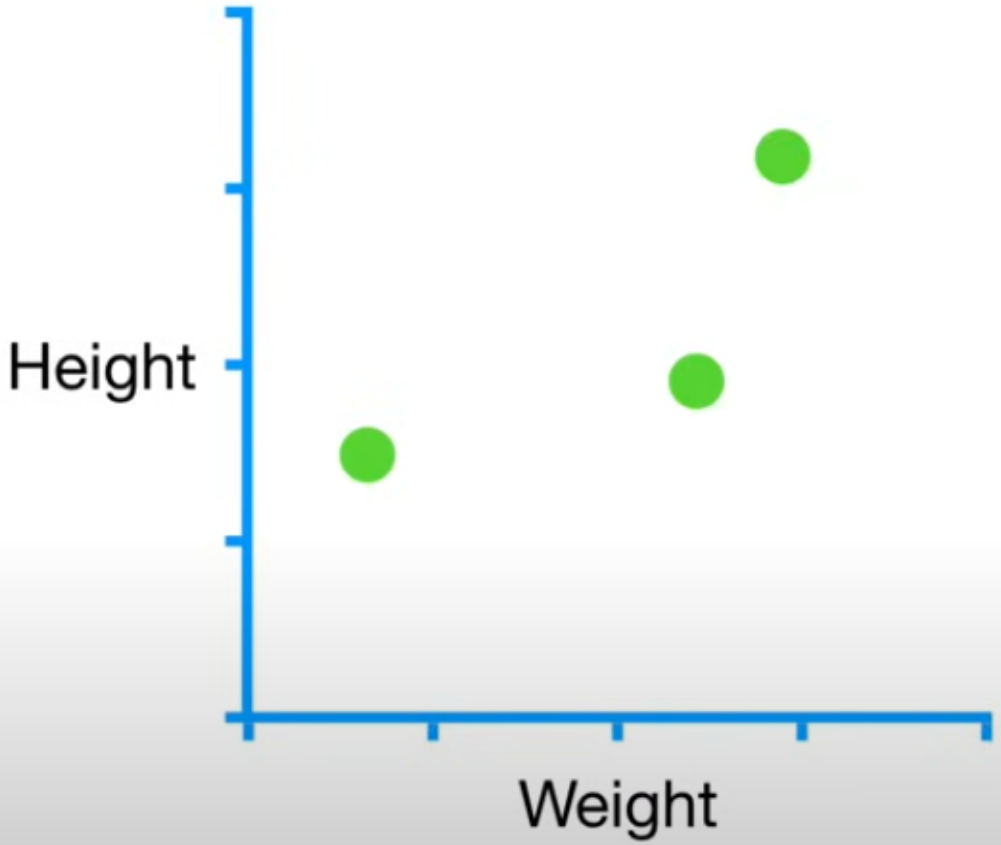
Predicted Height = intercept + 0.64 × **Weight**



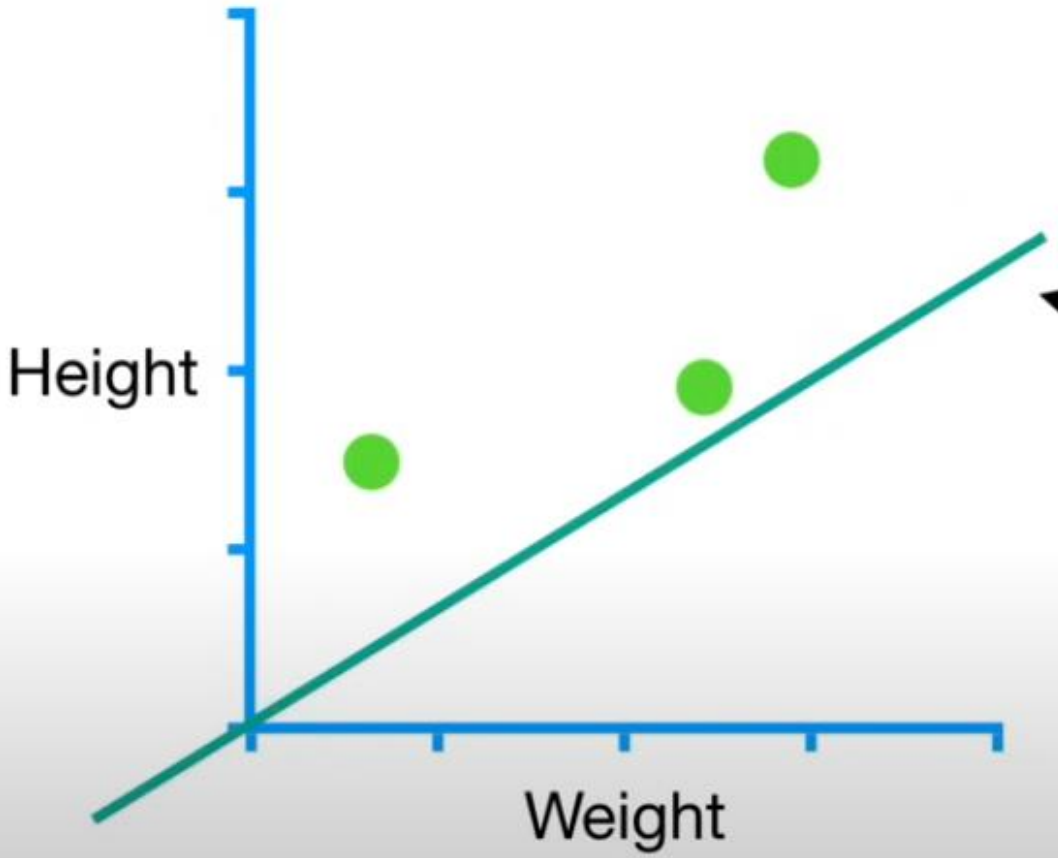
The first thing we do is pick a random value for the **Intercept**.

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

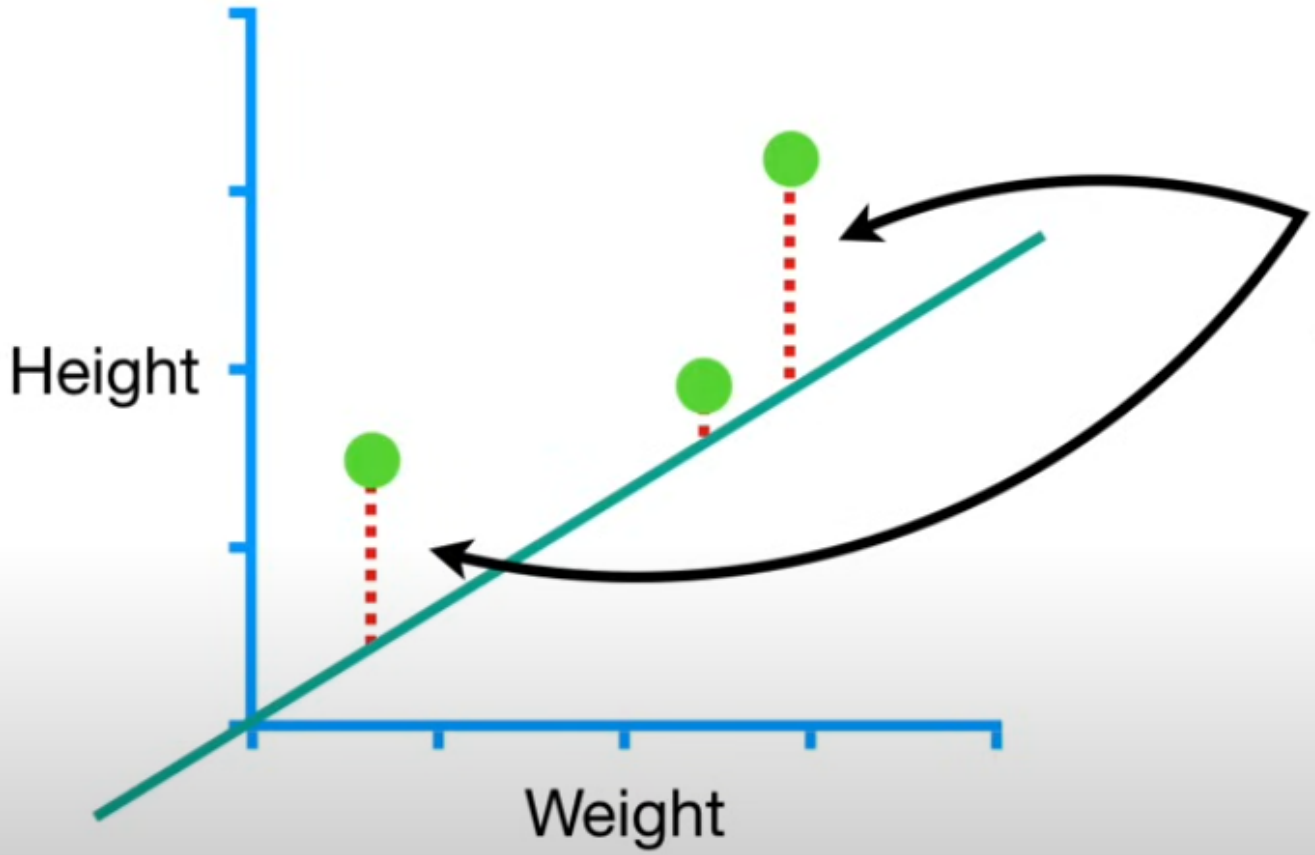
In this case, we'll use **0**,
but any number will do.



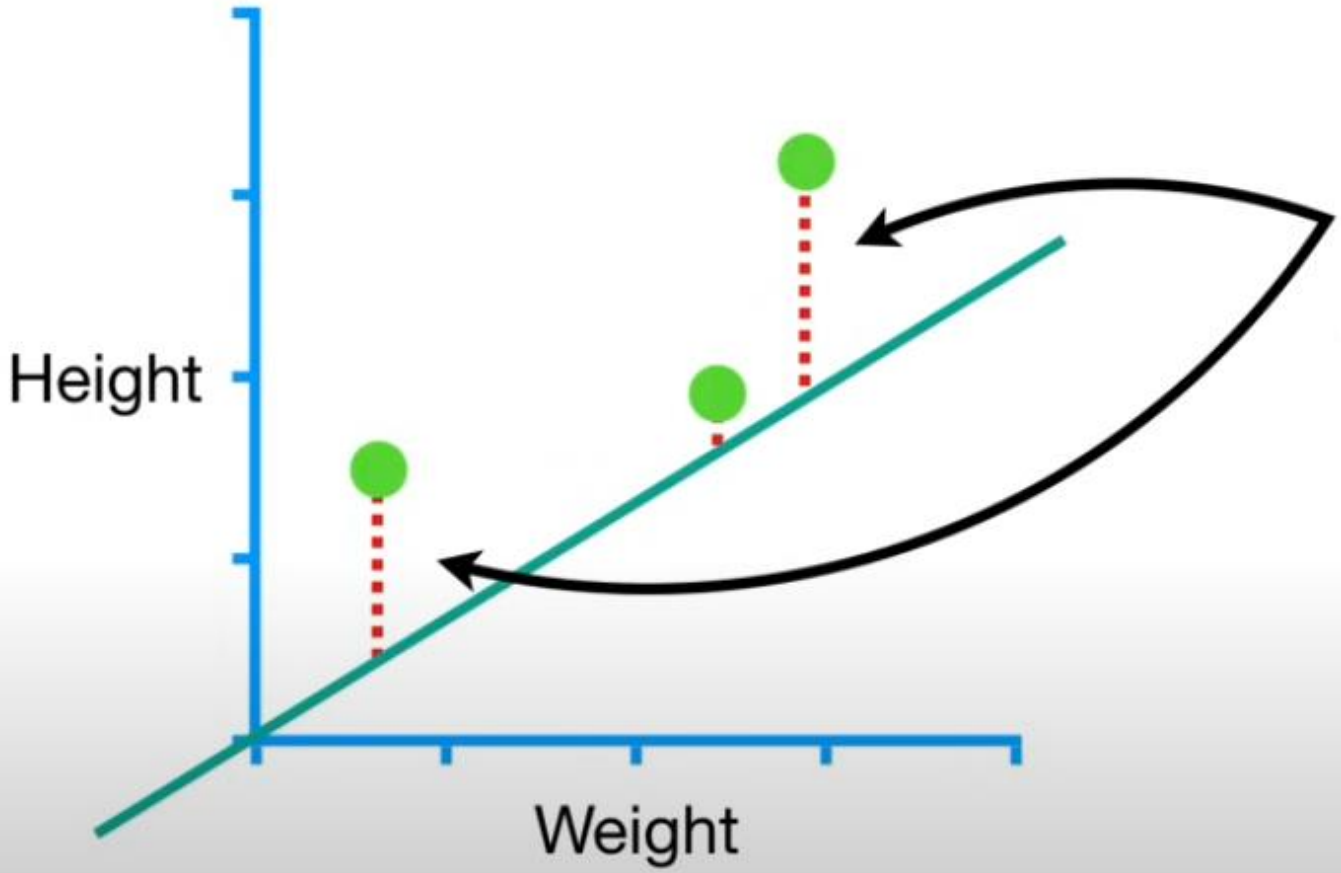
Predicted Height = 0 + 0.64 × **Weight**



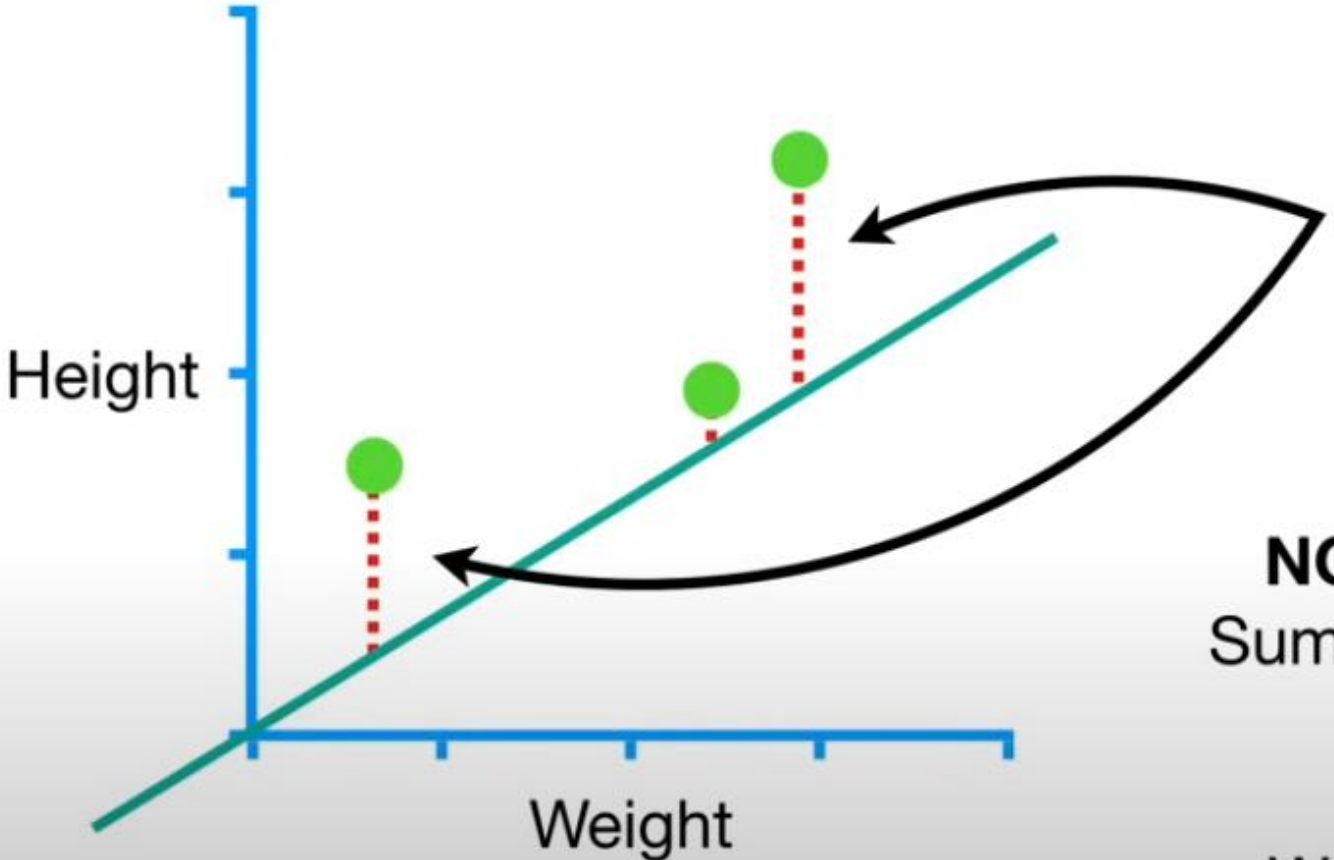
And that gives us the equation for this line.



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



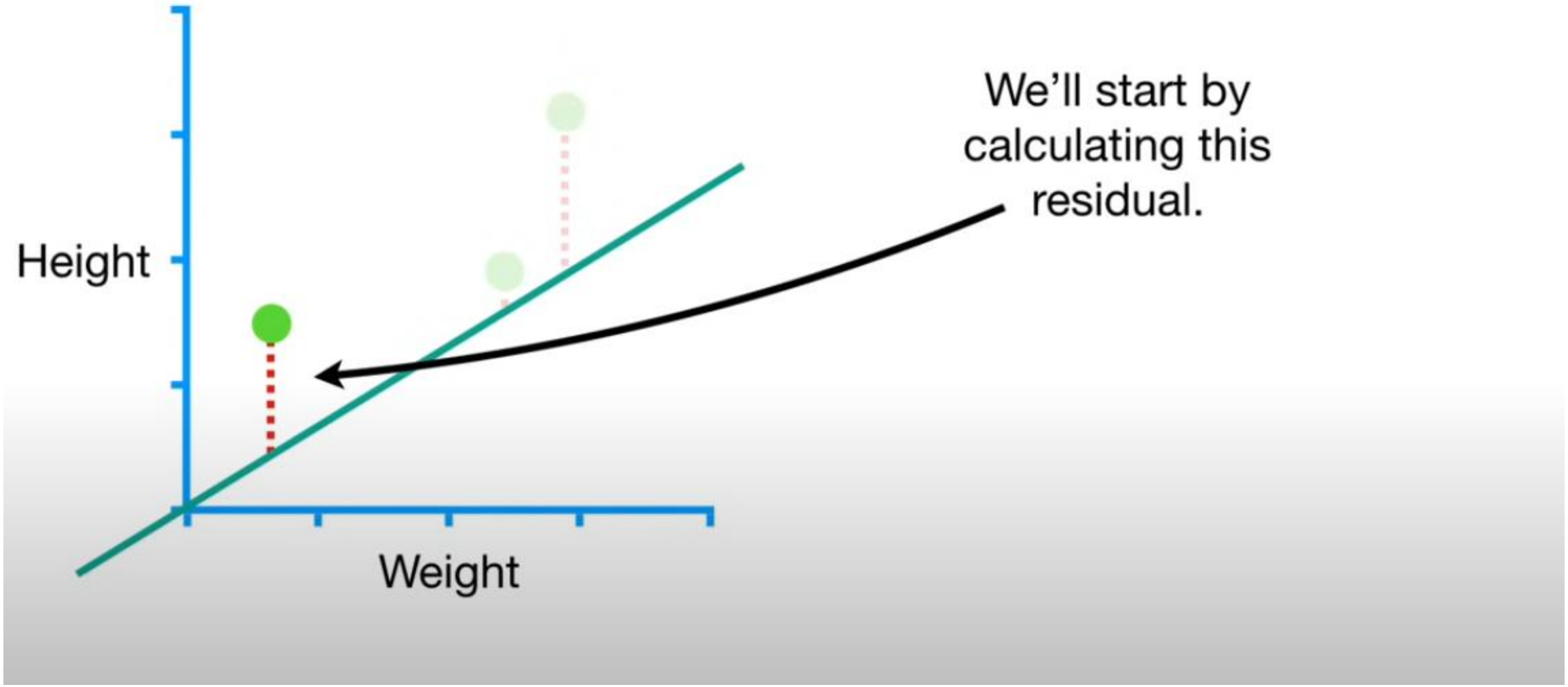
In this example, we will evaluate how well this line fits the data with the **Sum of the Squared Residuals**.

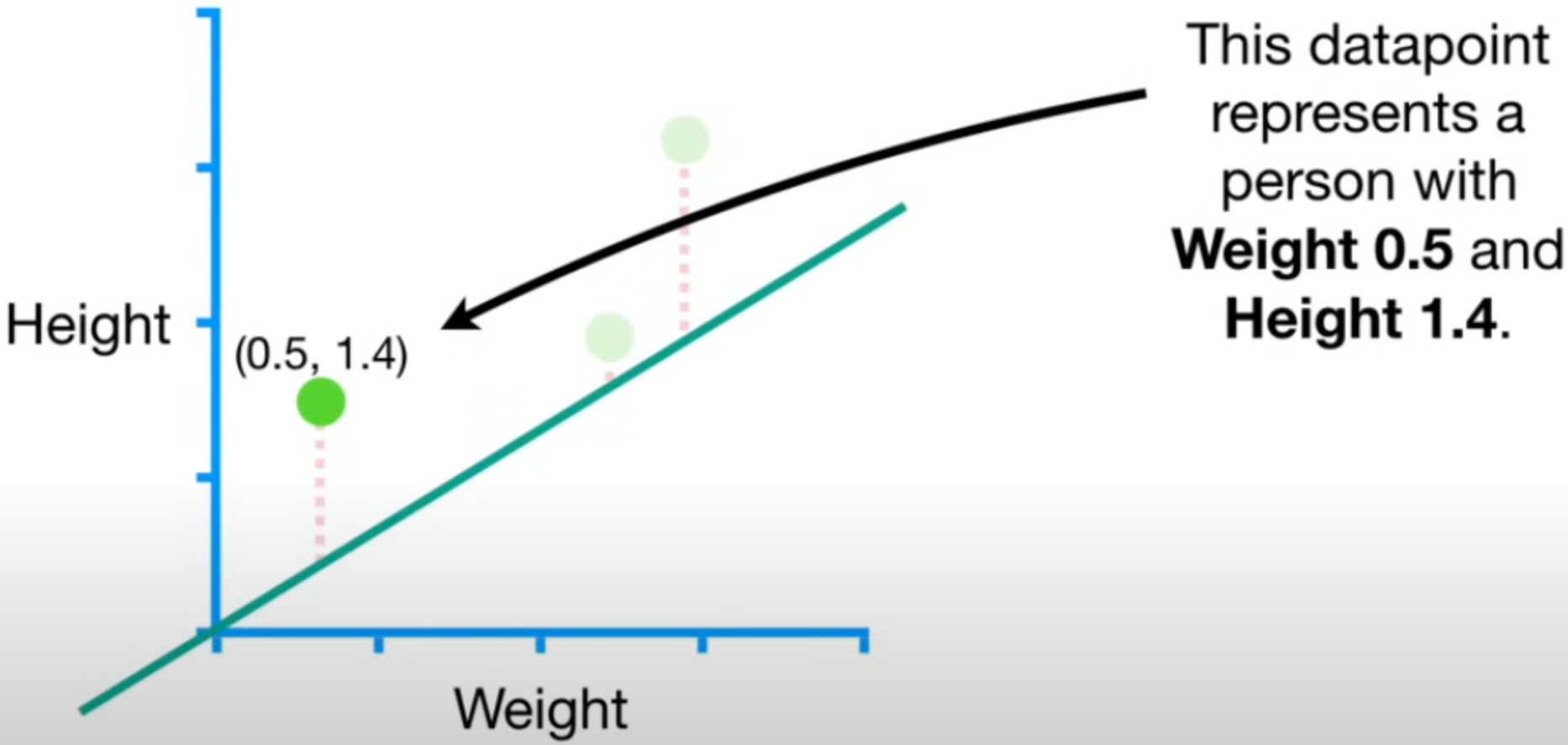


In this example, we will evaluate how well this line fits the data with the **Sum of the Squared Residuals**.

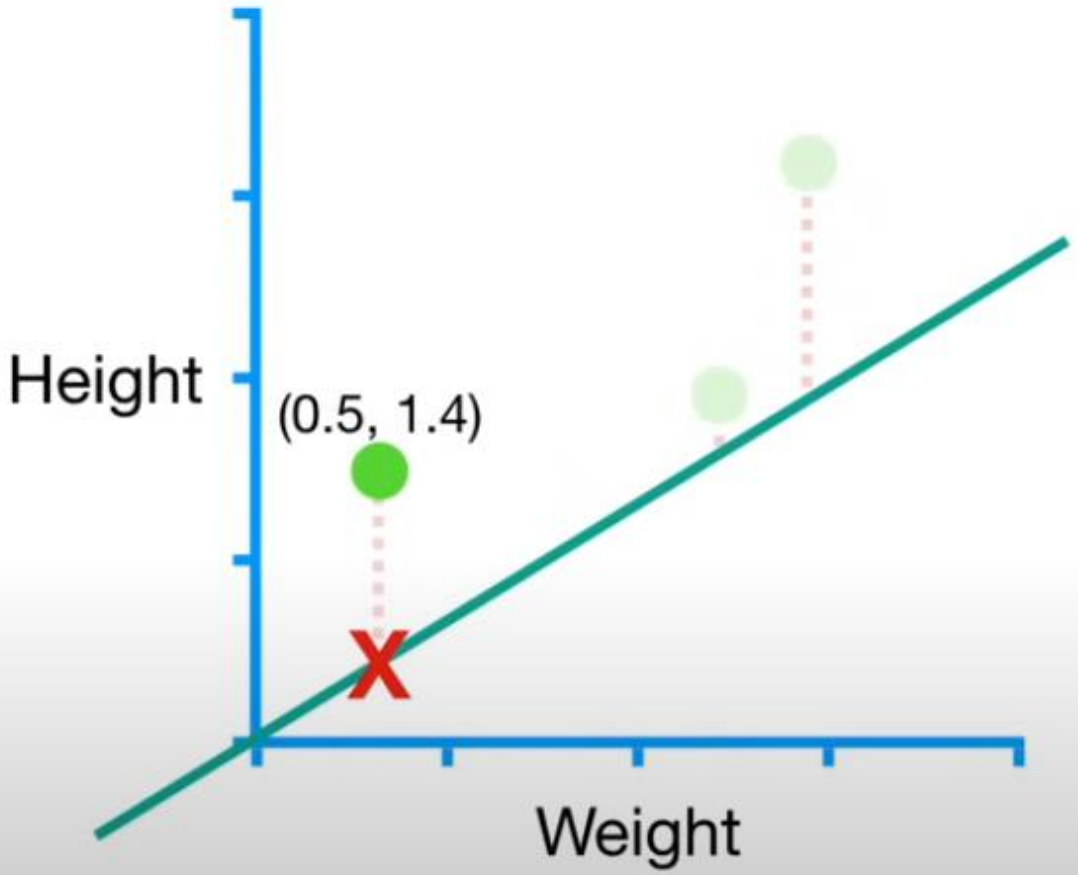
NOTE: In Machine Learning lingo, The Sum of the Squared Residuals is a type of **Loss Function**.

We'll talk more about **Loss Functions** towards the end of the video.


We'll start by calculating this residual.
Height
Weight



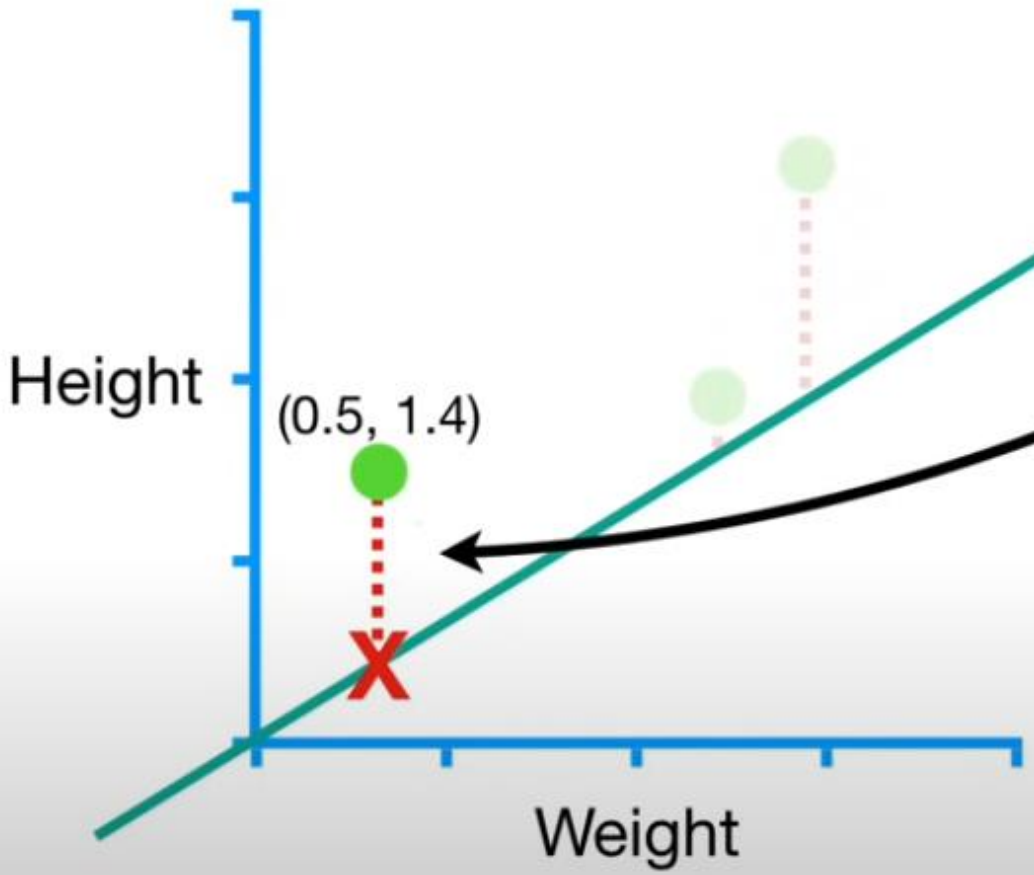
This datapoint represents a person with **Weight 0.5** and **Height 1.4**.



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

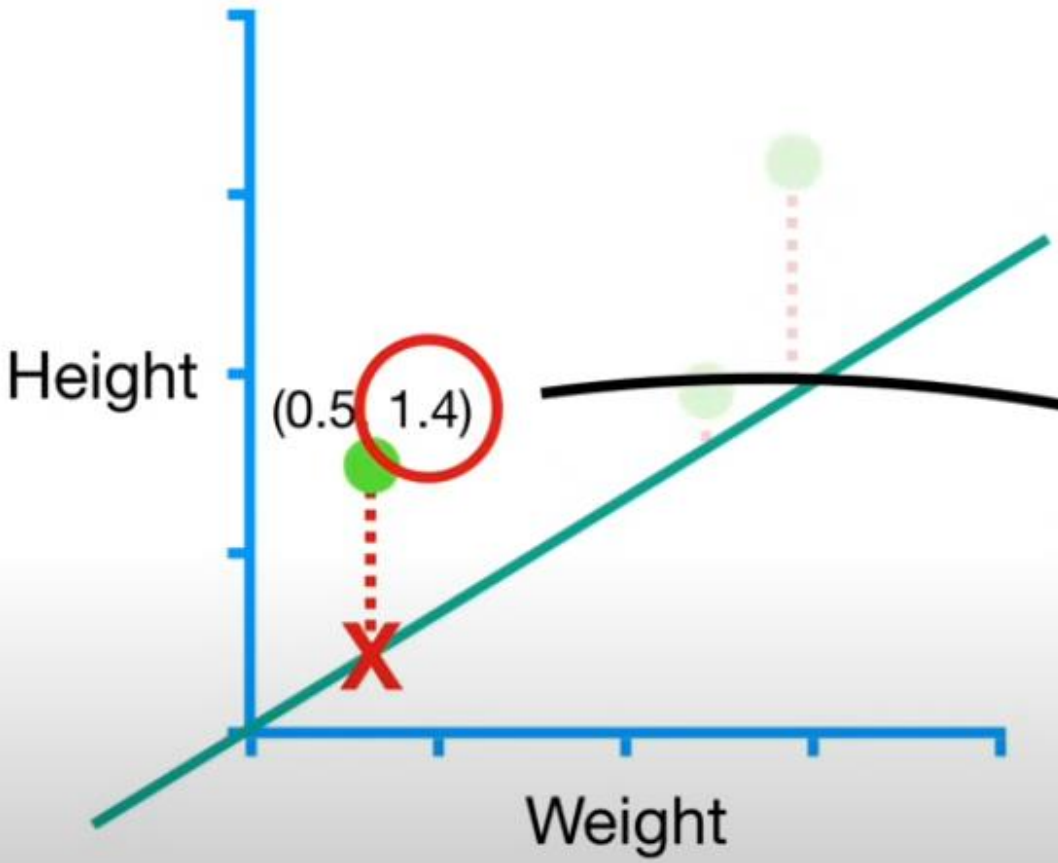
…by plugging **Weight = 0.5** into the equation for the line…

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



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

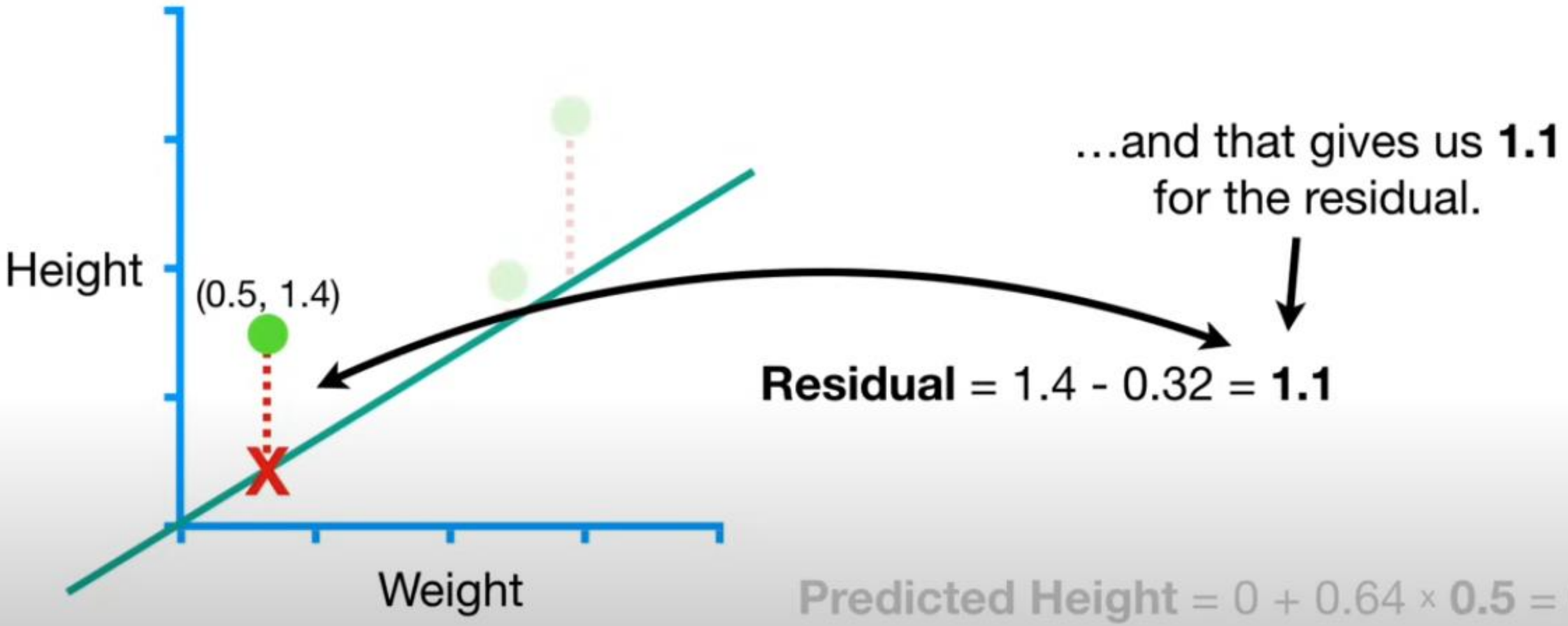
Predicted Height = 0 + 0.64 × **0.5** = **0.32**



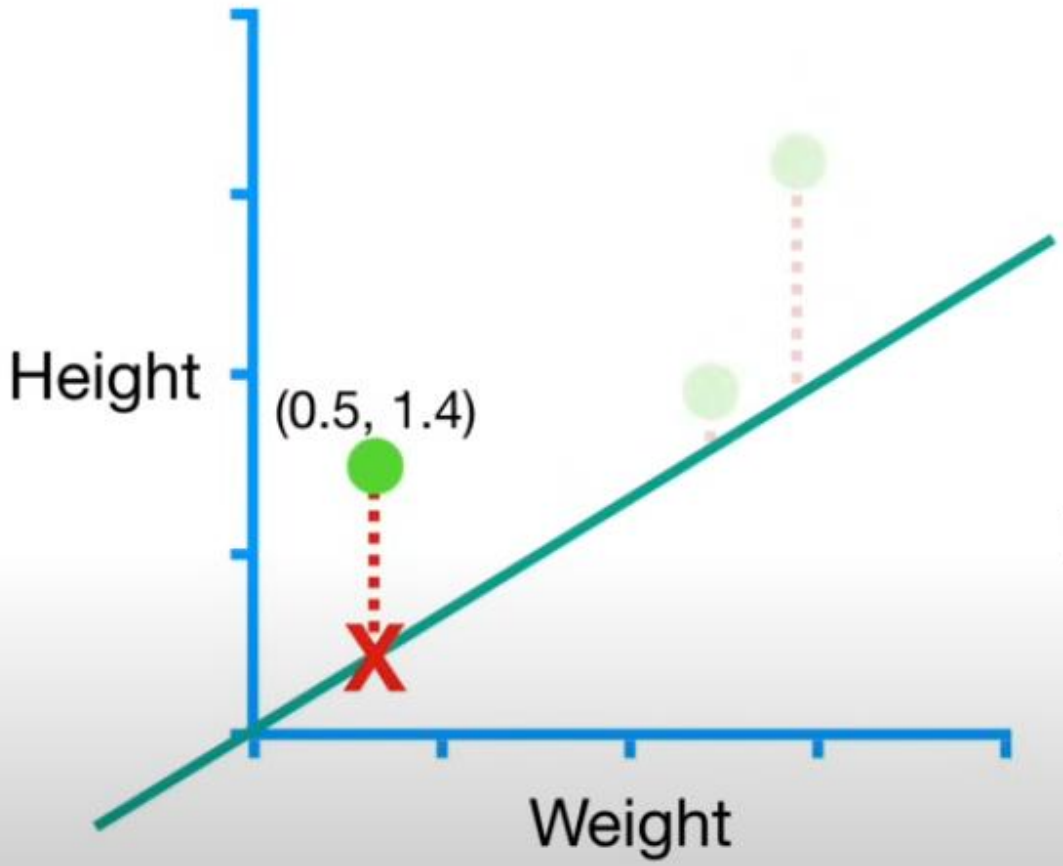
…so we calculate the difference between **1.4** (the **Observed Height**)…

Residual = 1.4

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


Height
(0.5, 1.4)
Weight
…and that gives us **1.1**
for the residual.
Residual = 1.4 - 0.32 = **1.1**
Predicted Height = 0 + 0.64 × **0.5** = **0.32**

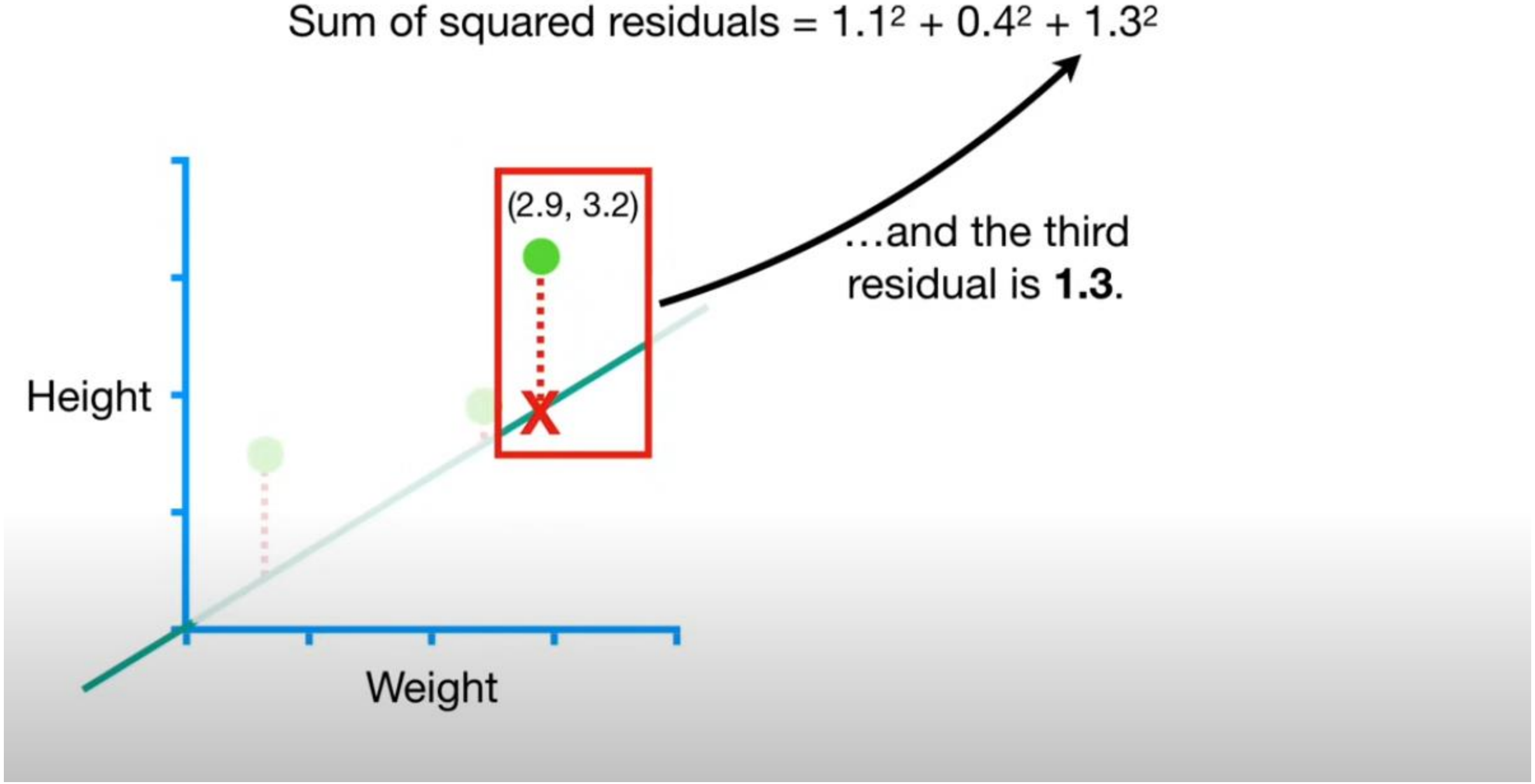
Sum of squared residuals = 1.1^2



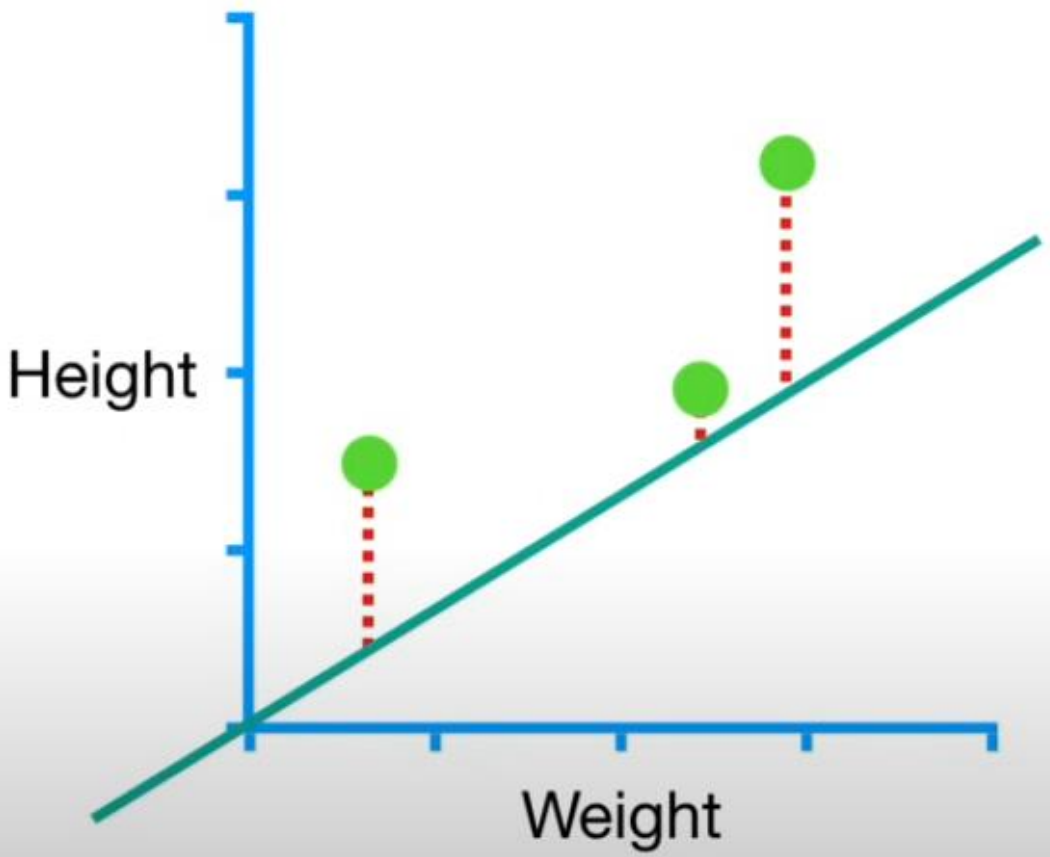
Here's the square of the first residual…

Residual = 1.4 - 0.32 = **1.1**

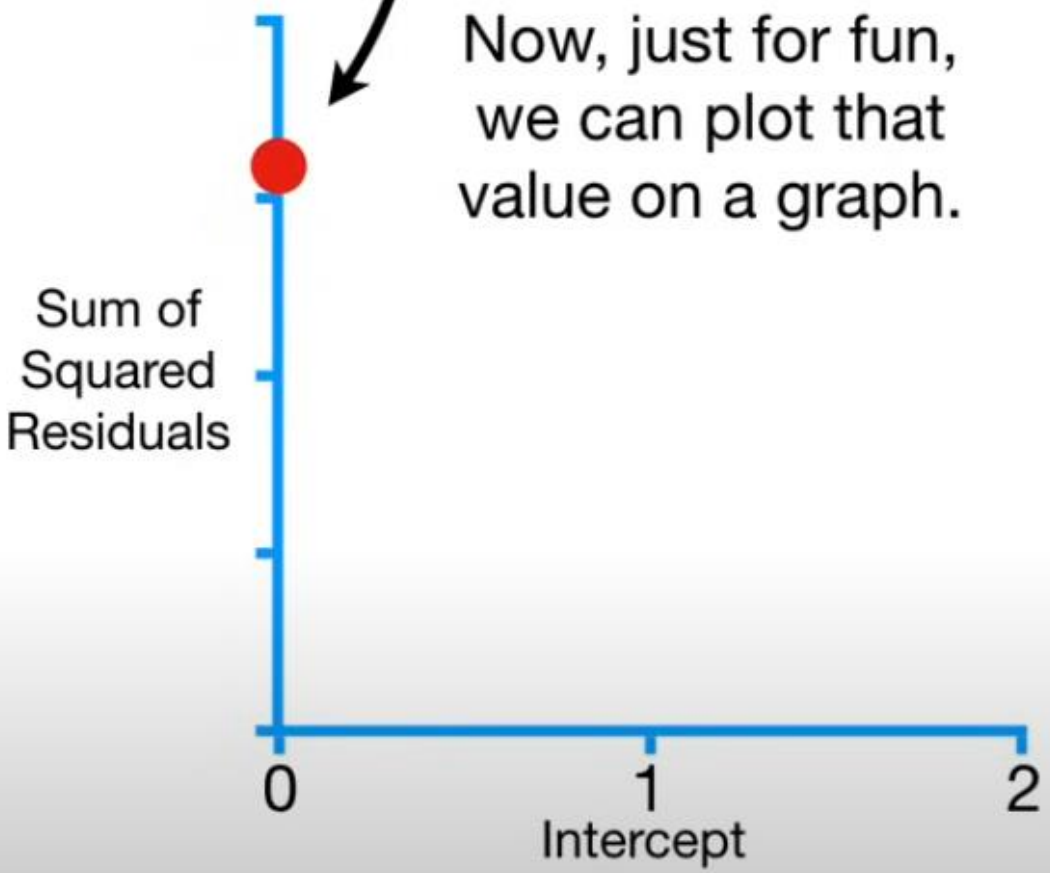
Predicted Height = 0 + 0.64 × **0.5** = **0.32**


Sum of squared residuals = $1.1^2 + 0.4^2 + 1.3^2$
(2.9, 3.2)
…and the third residual is **1.3**.
Height
Weight

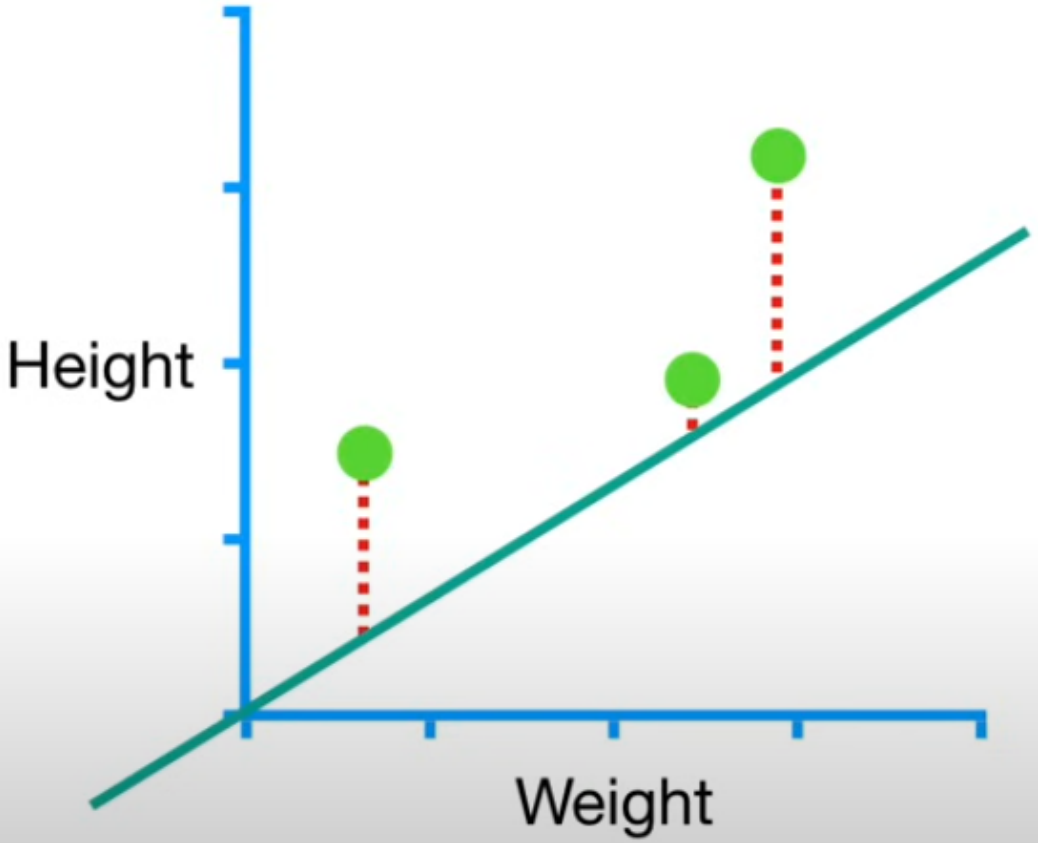
Sum of squared residuals = $1.1^2 + 0.4^2 + 1.3^2 = 3.1$

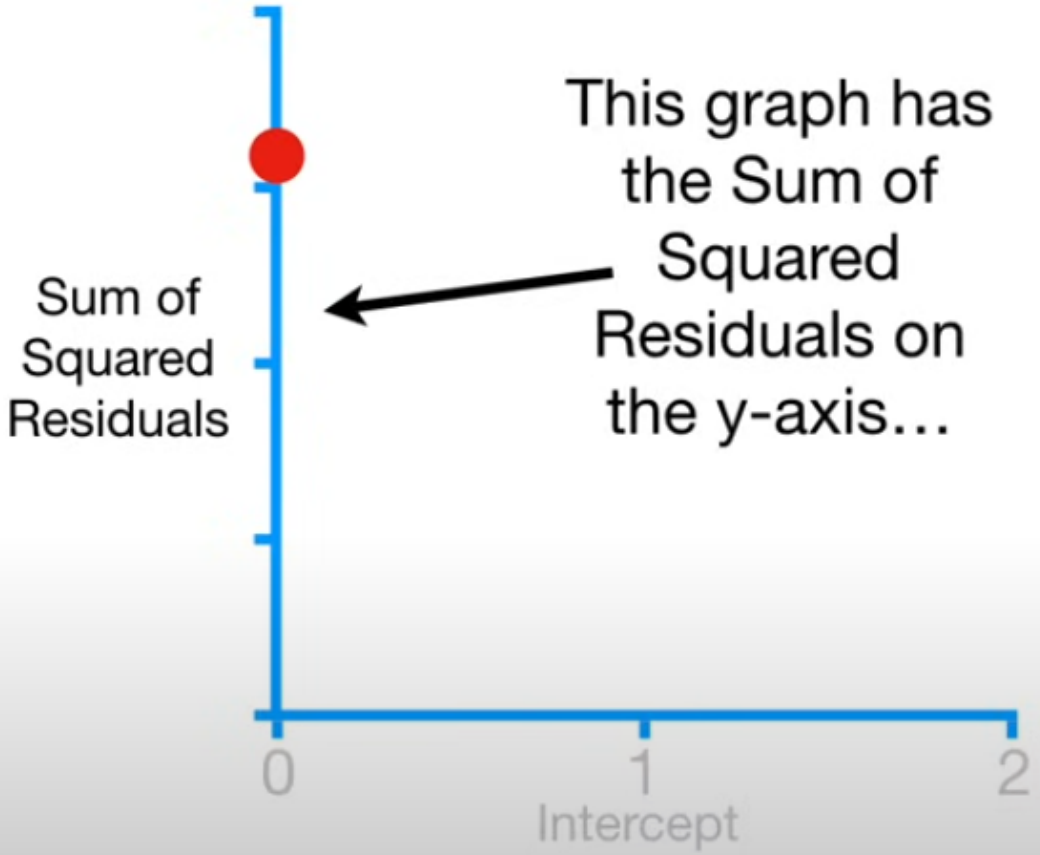


Now, just for fun, we can plot that value on a graph.

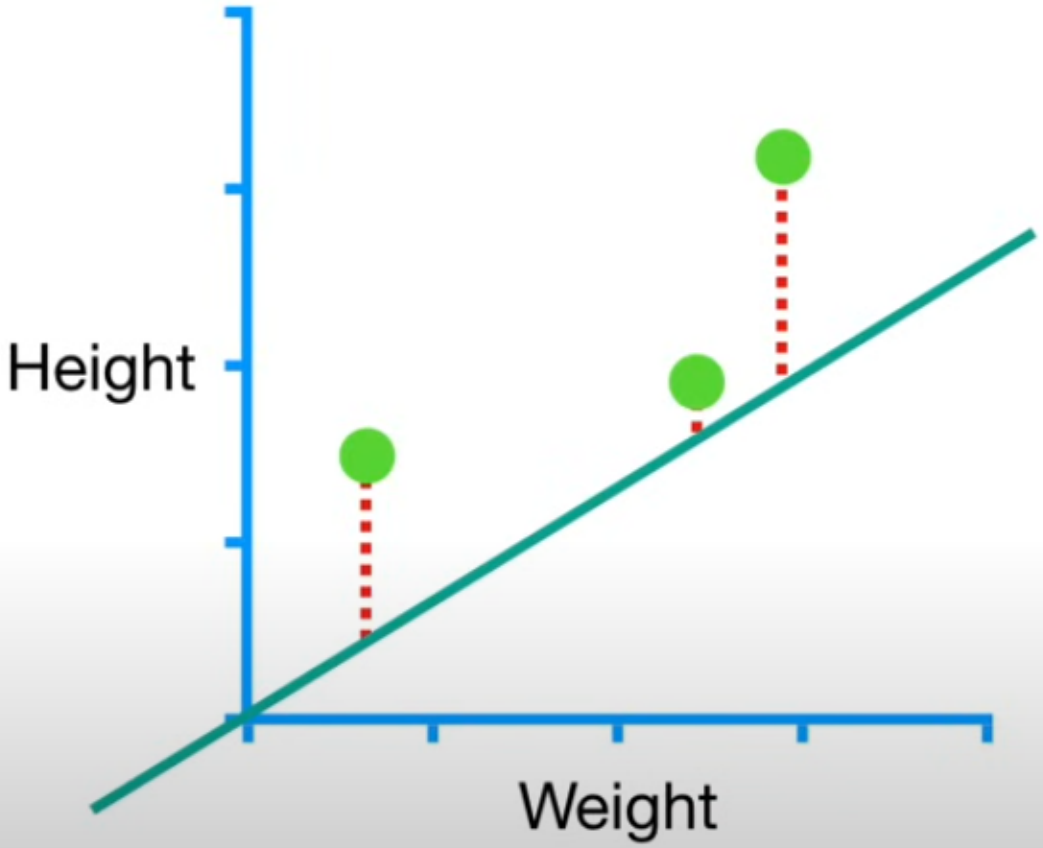


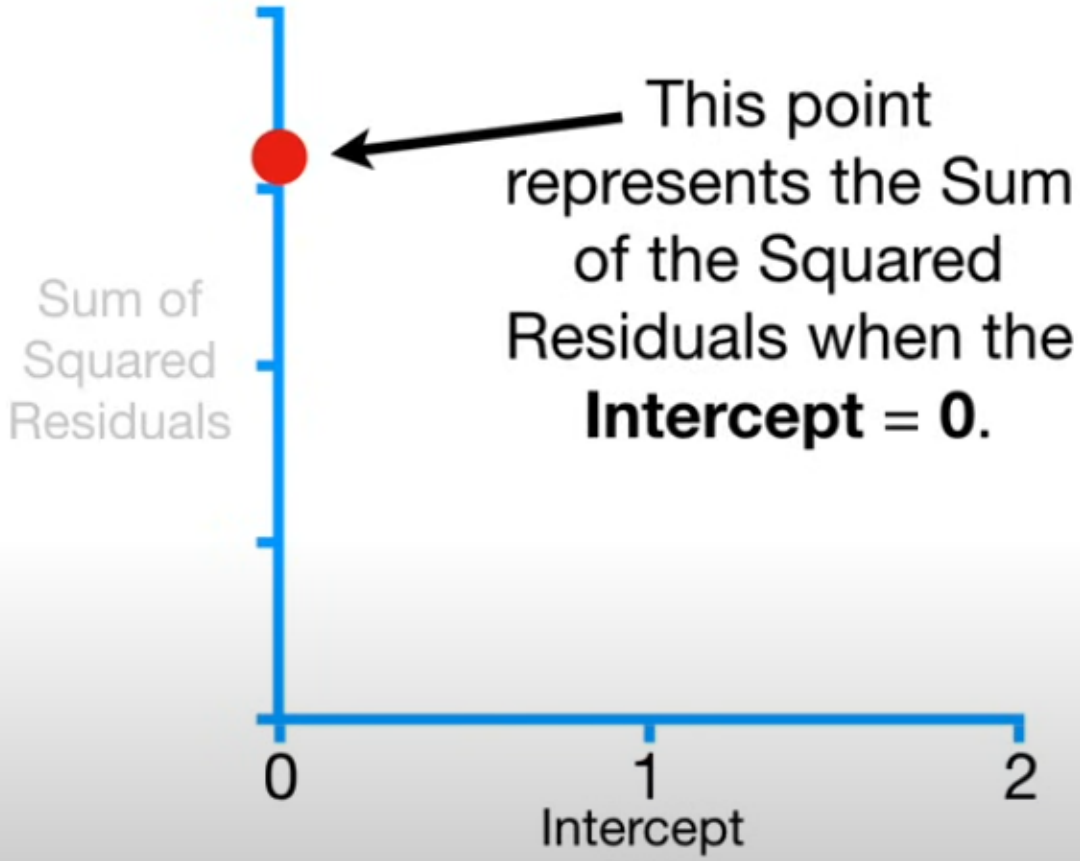
Sum of squared residuals = $1.1^2 + 0.4^2 + 1.3^2 = 3.1$



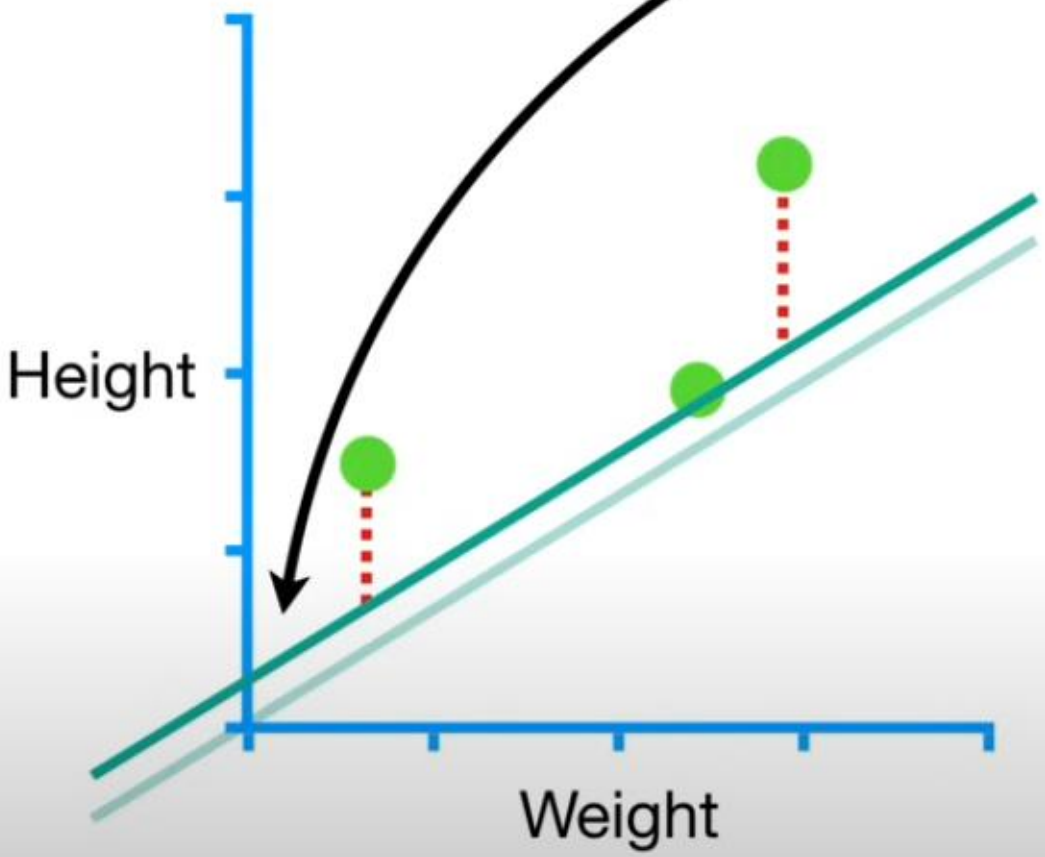


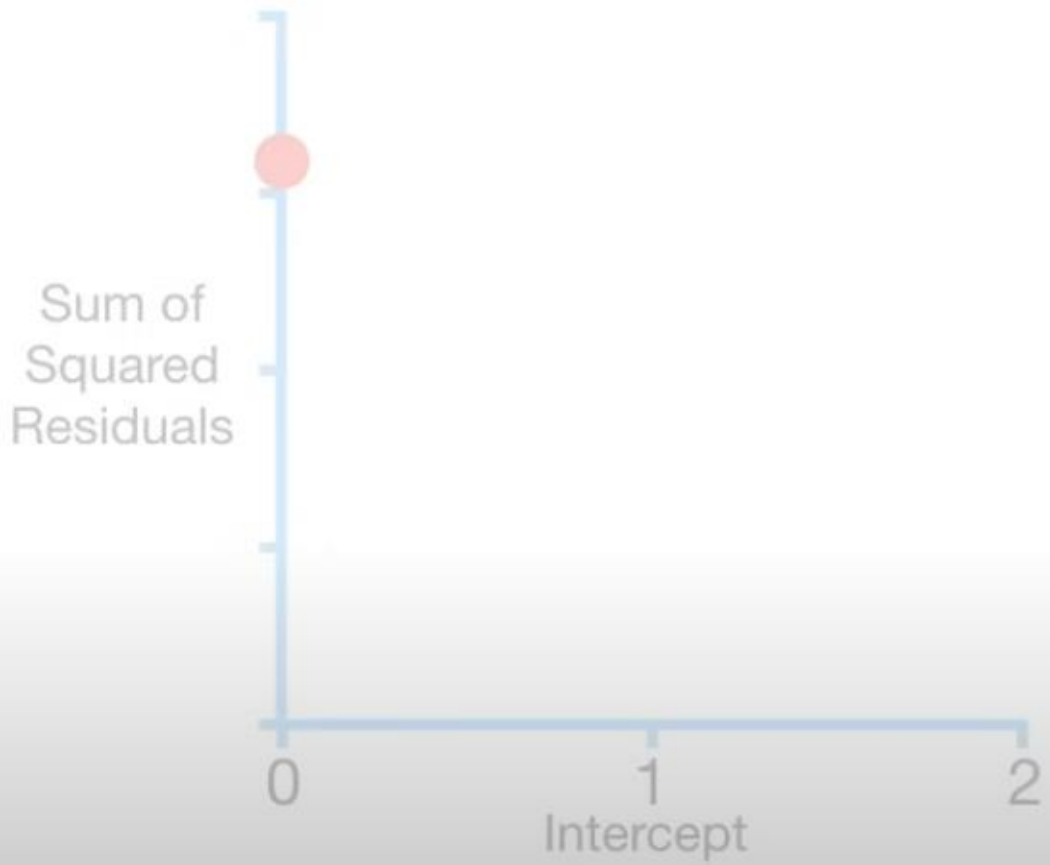
Sum of squared residuals = $1.1^2 + 0.4^2 + 1.3^2 = 3.1$



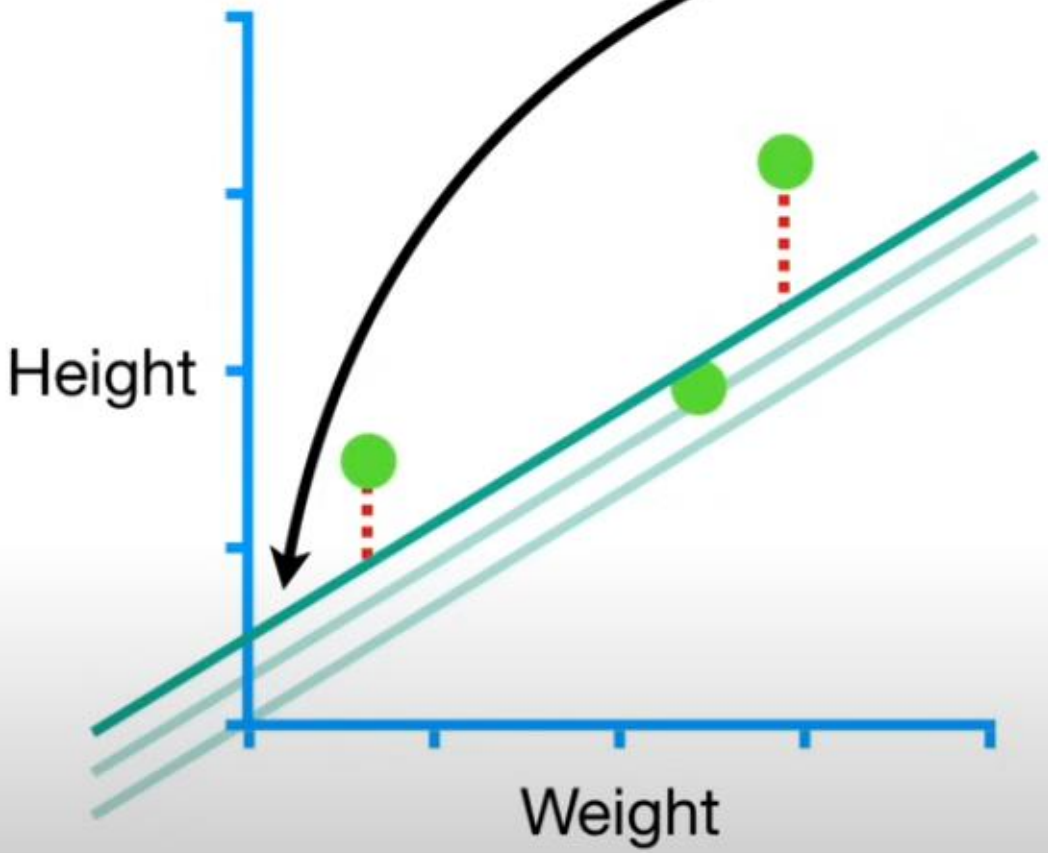


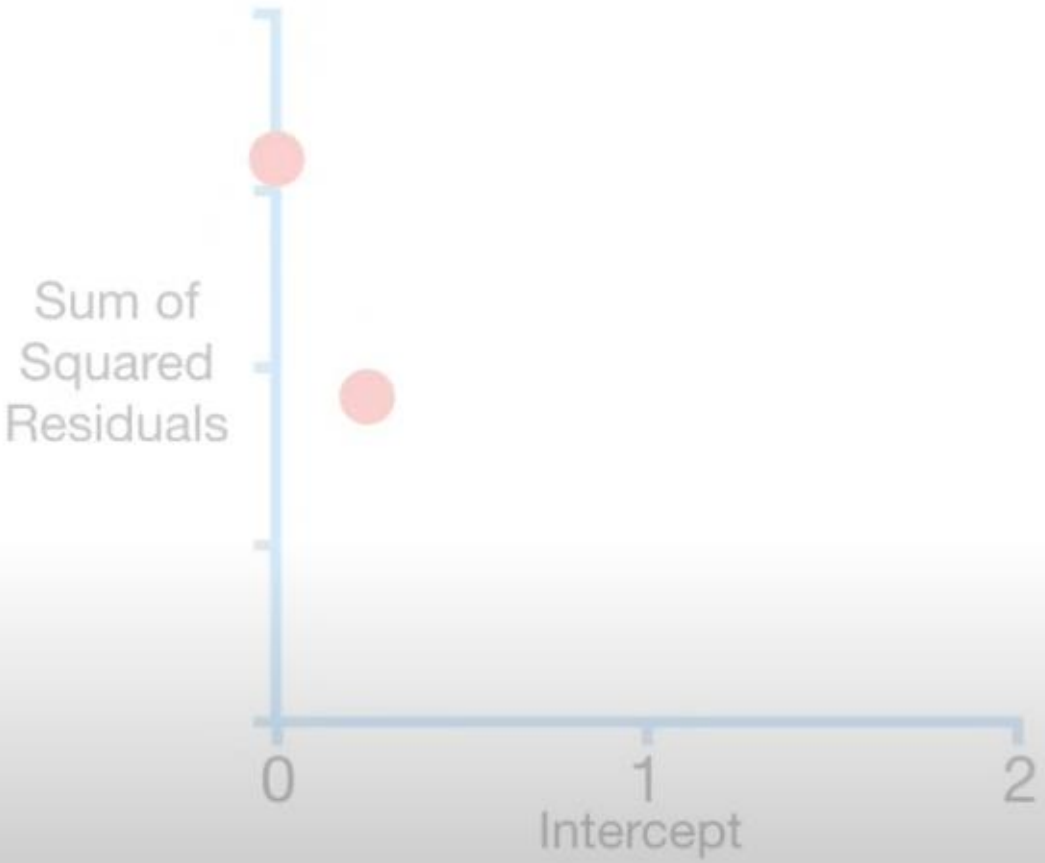
However, if the
Intercept = 0.25…



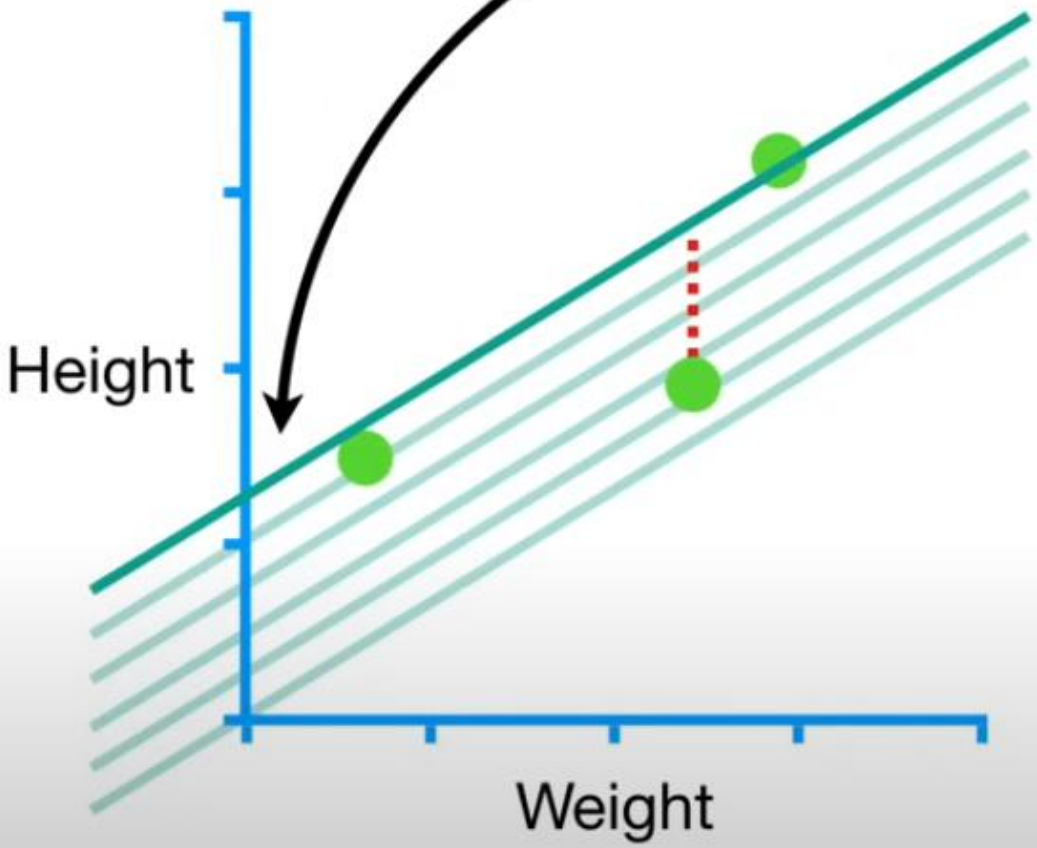


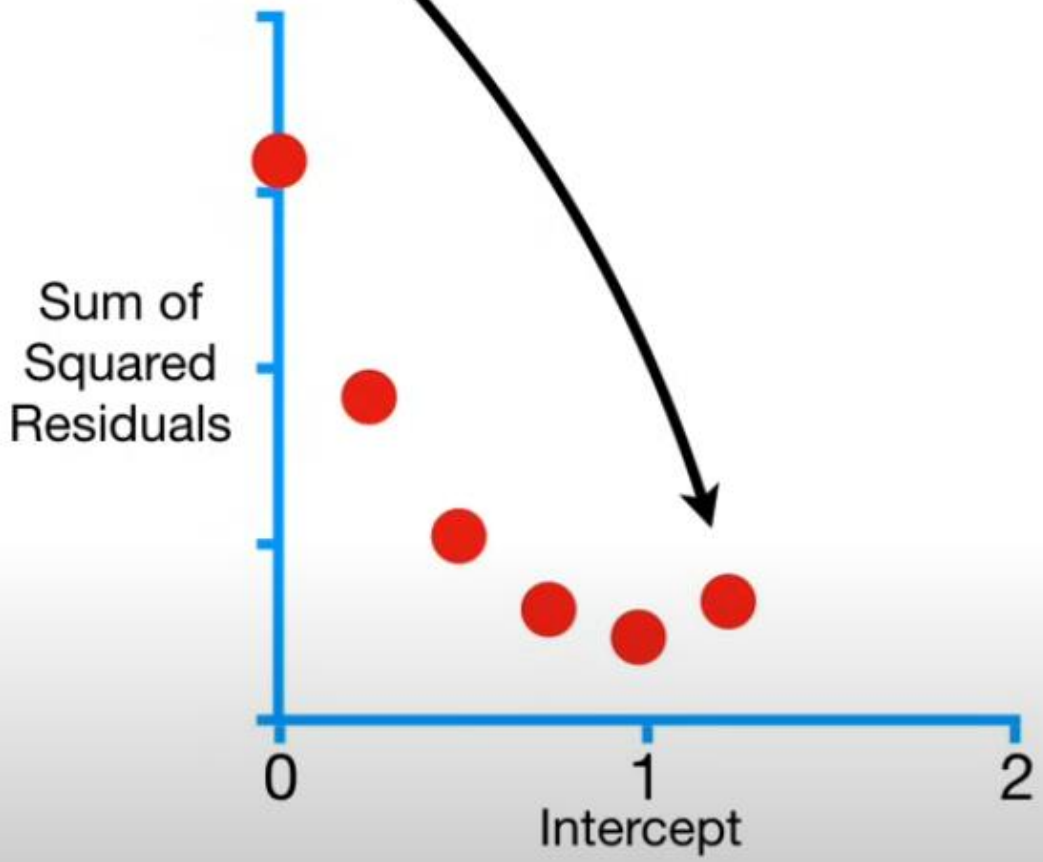
And if the
Intercept = 0.5…



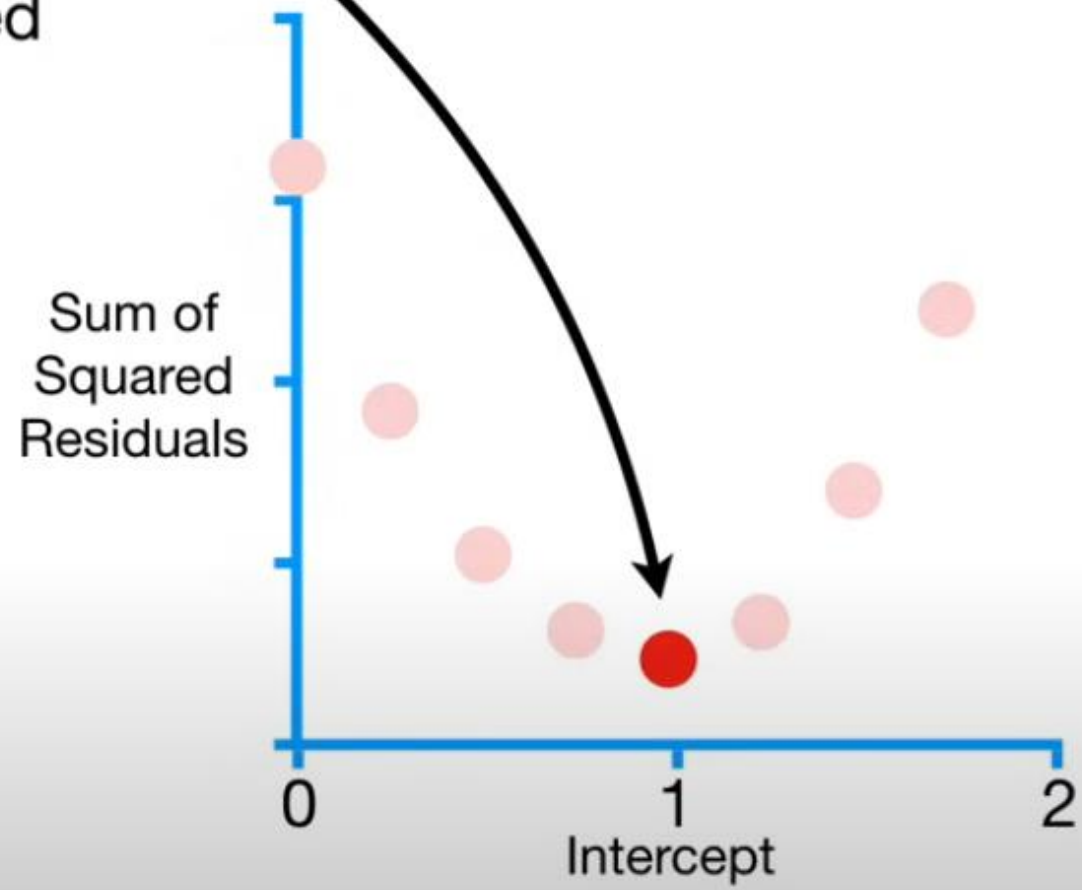


And for increasing values for the **Intercept**, we get these points.

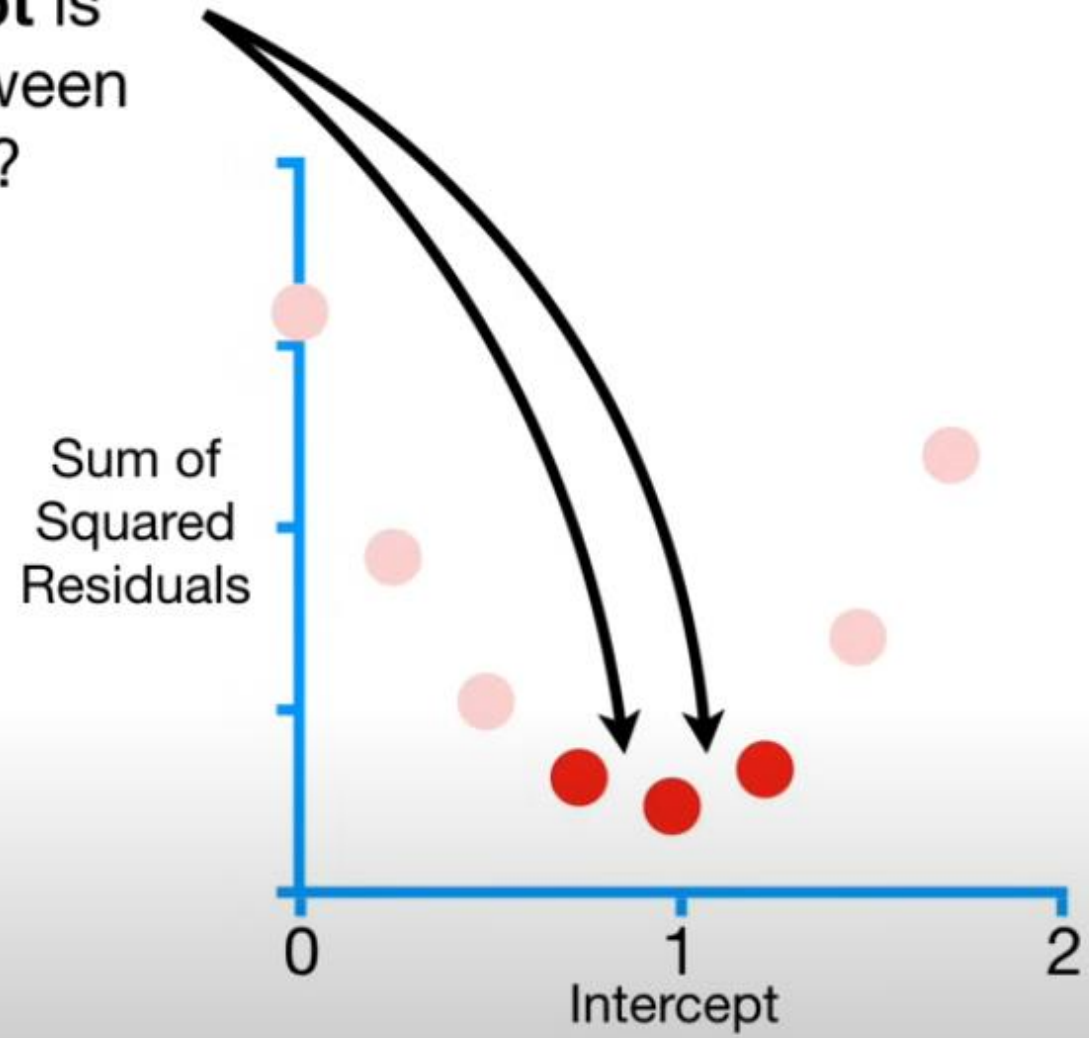




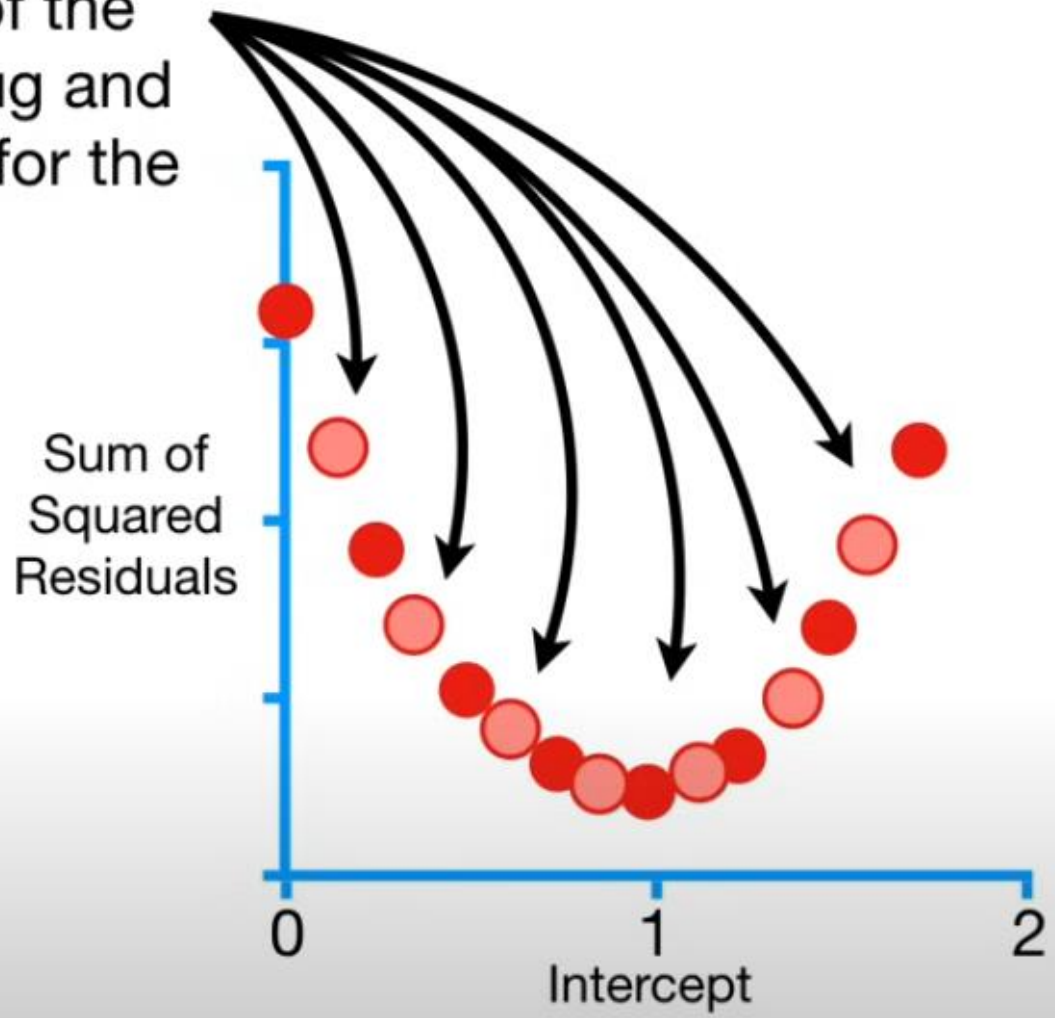
Of the points that we calculated for the graph, this one has the lowest Sum of Squared Residuals…



What if the best value for the **Intercept** is somewhere between these values?

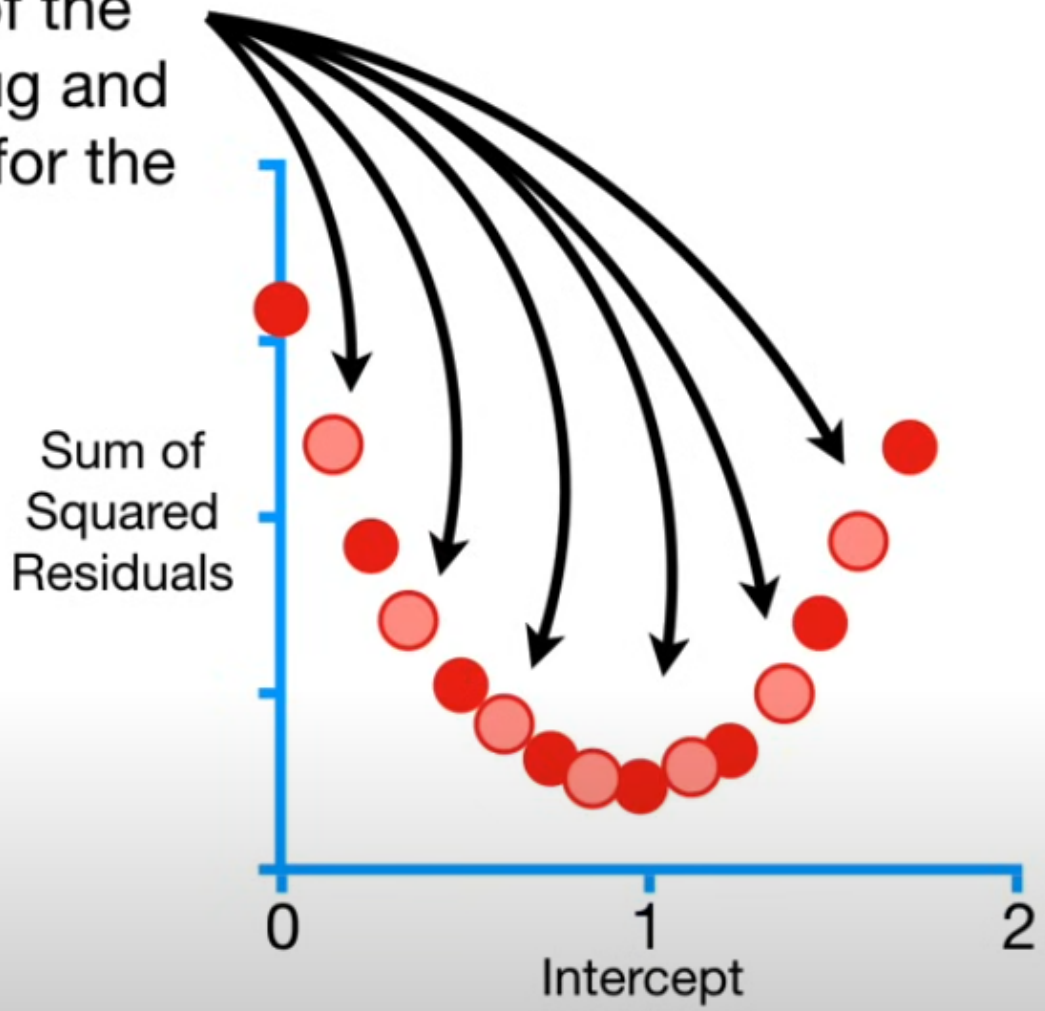


A slow and painful method for finding the minimal Sum of the Squared Residuals is to plug and chug a bunch more values for the **Intercept**.



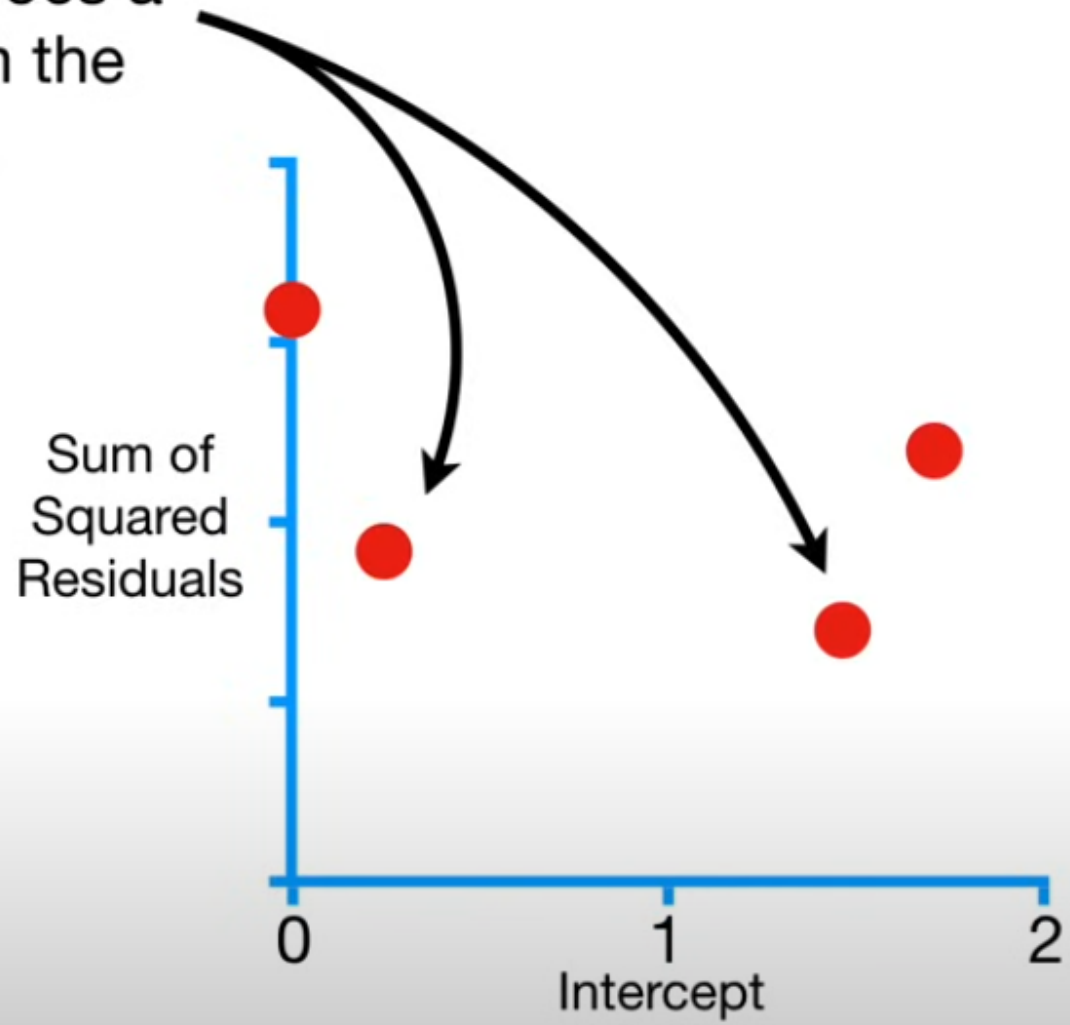
A slow and painful method for finding the minimal Sum of the Squared Residuals is to plug and chug a bunch more values for the **Intercept**.

Ugh.

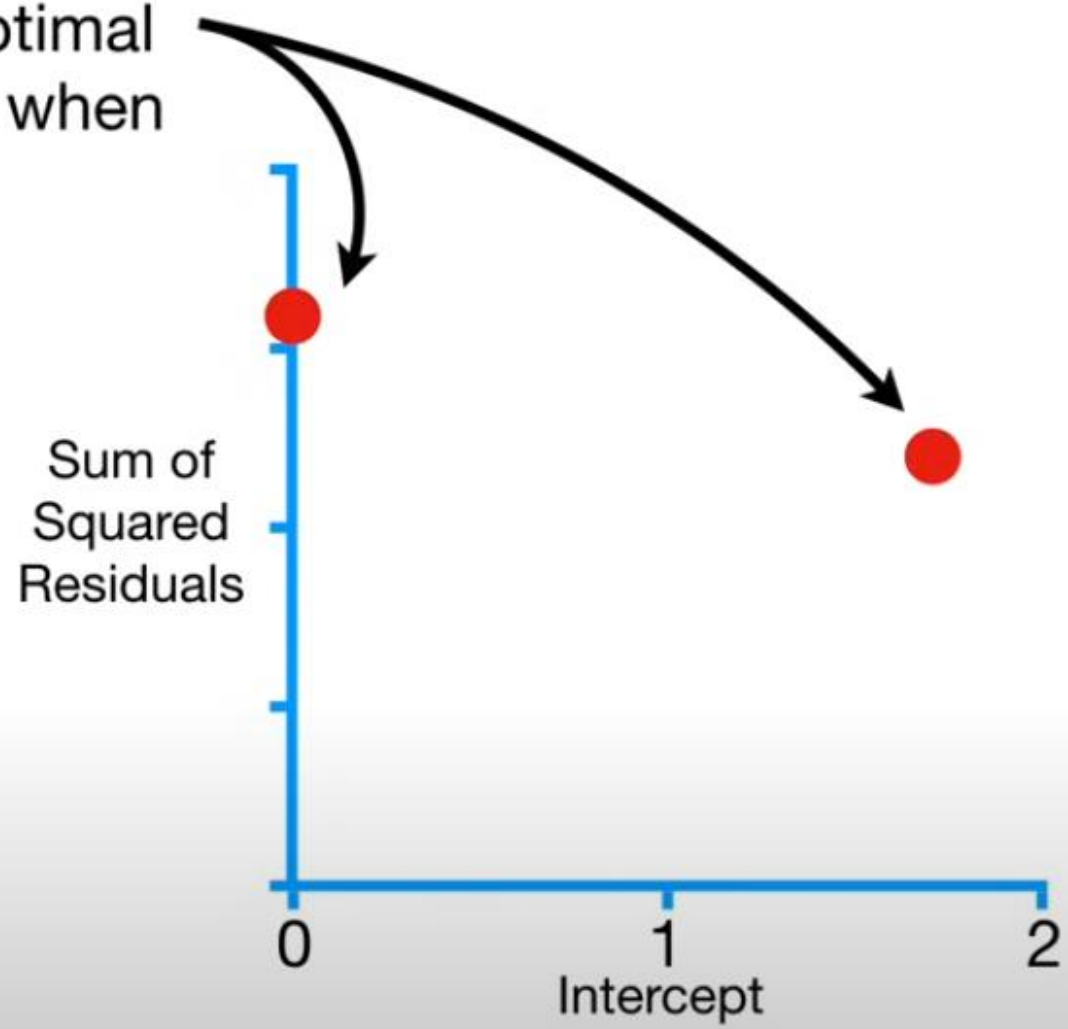


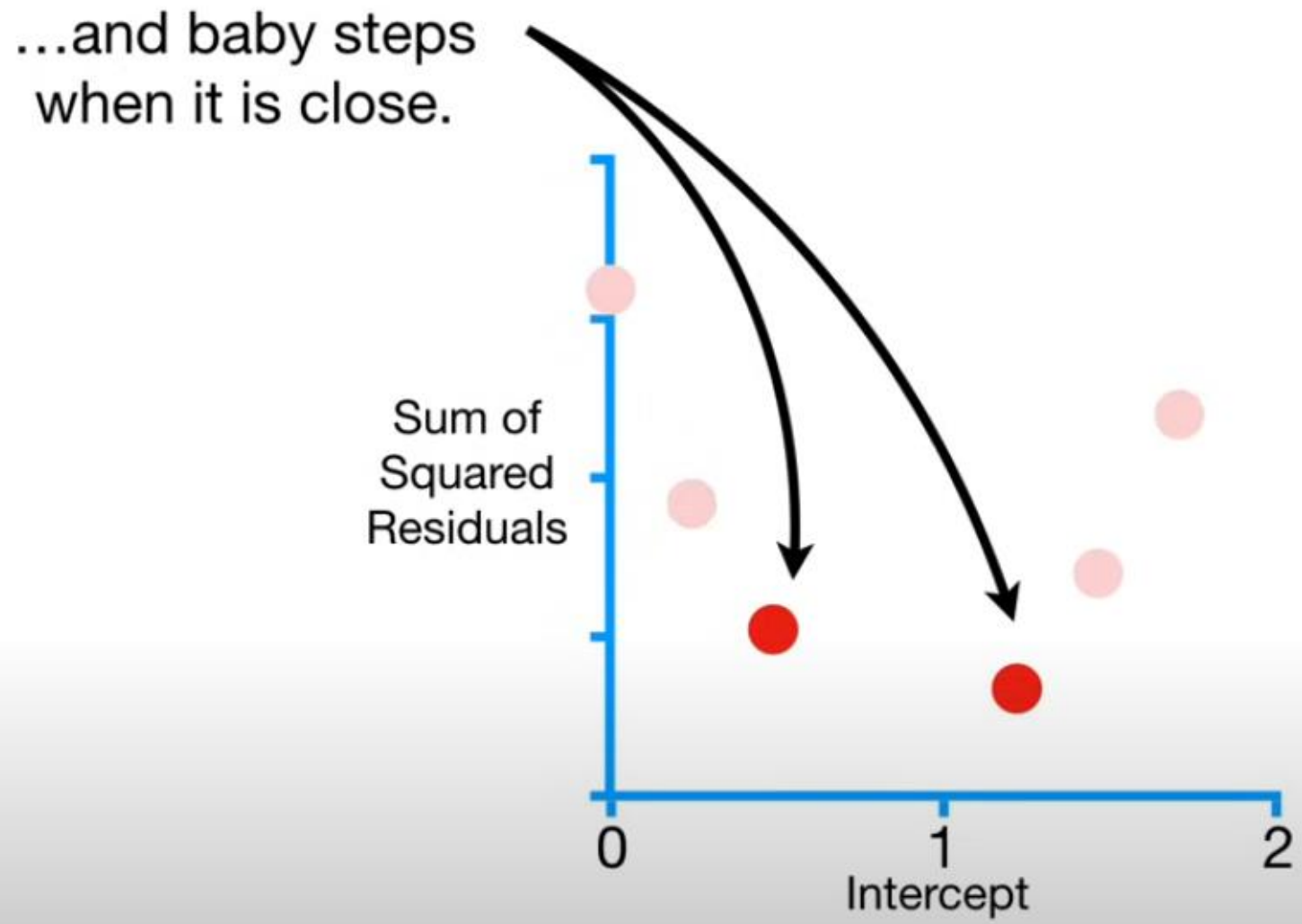
Don't despair!
Gradient Descent is way more efficient!

Gradient Descent only does a few calculations far from the optimal solution…

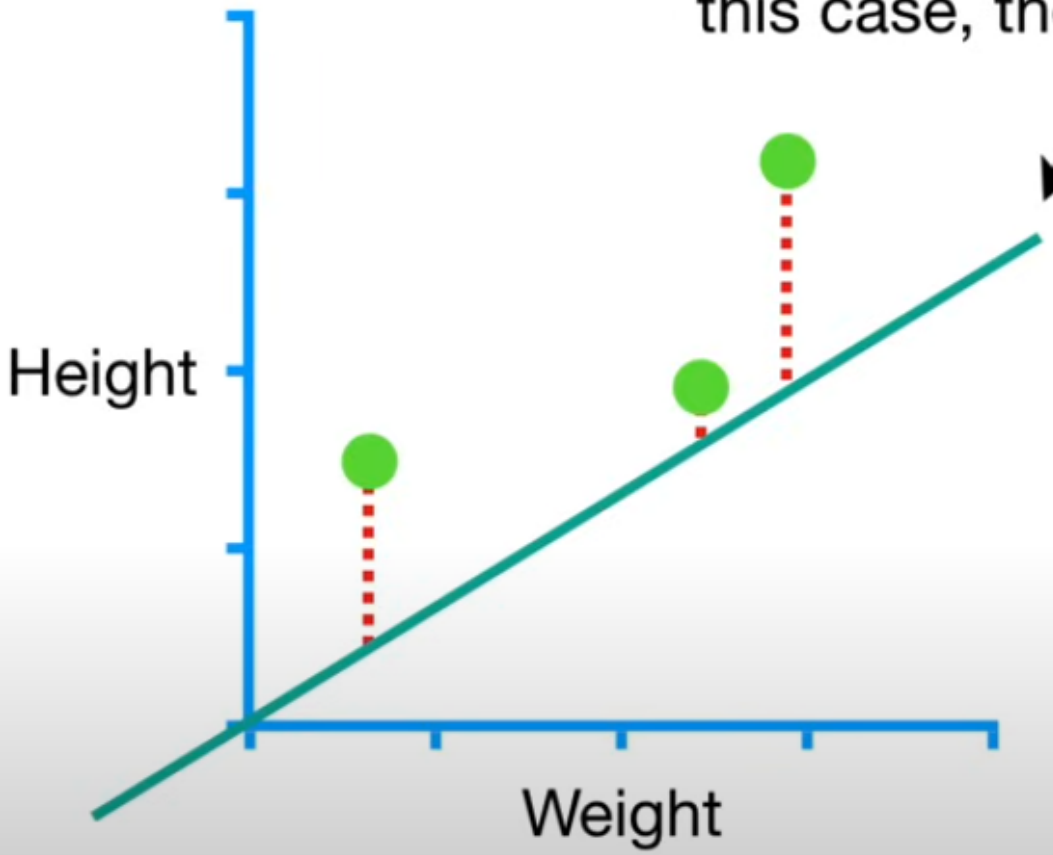


In other words, **Gradient Descent** identifies the optimal value by taking big steps when it is far away…



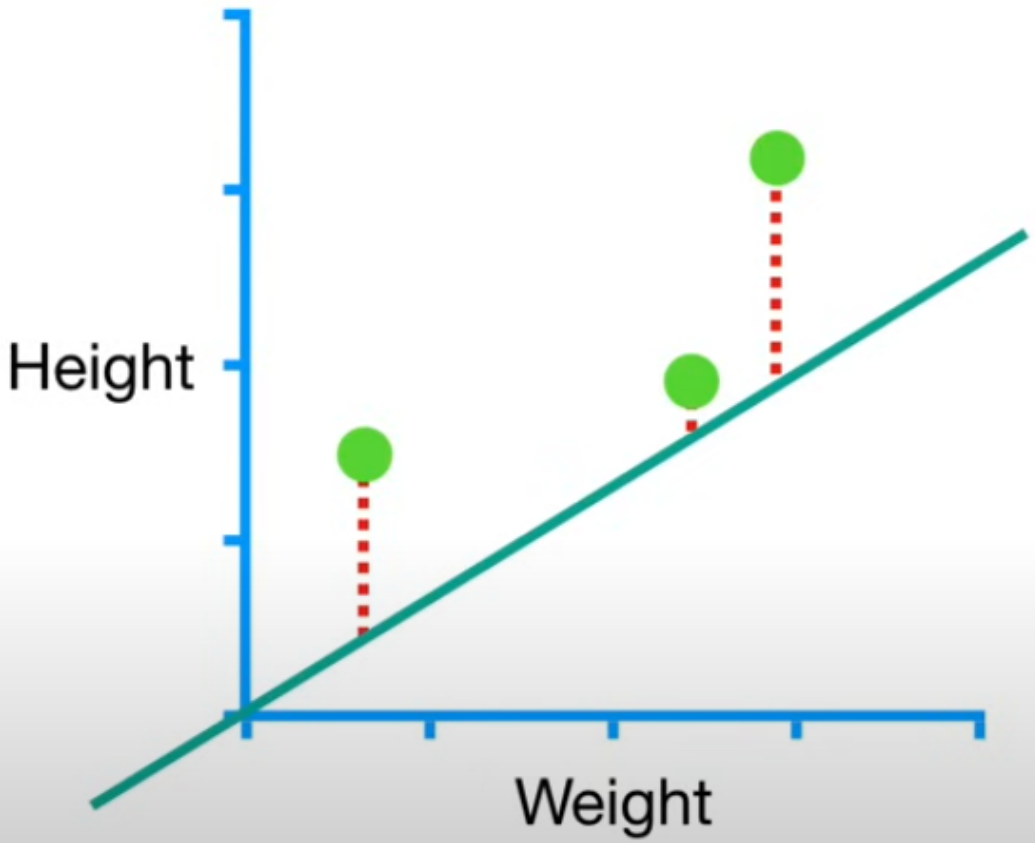

…and baby steps
when it is close.
Sum of
Squared
Residuals
0
1
2
Intercept

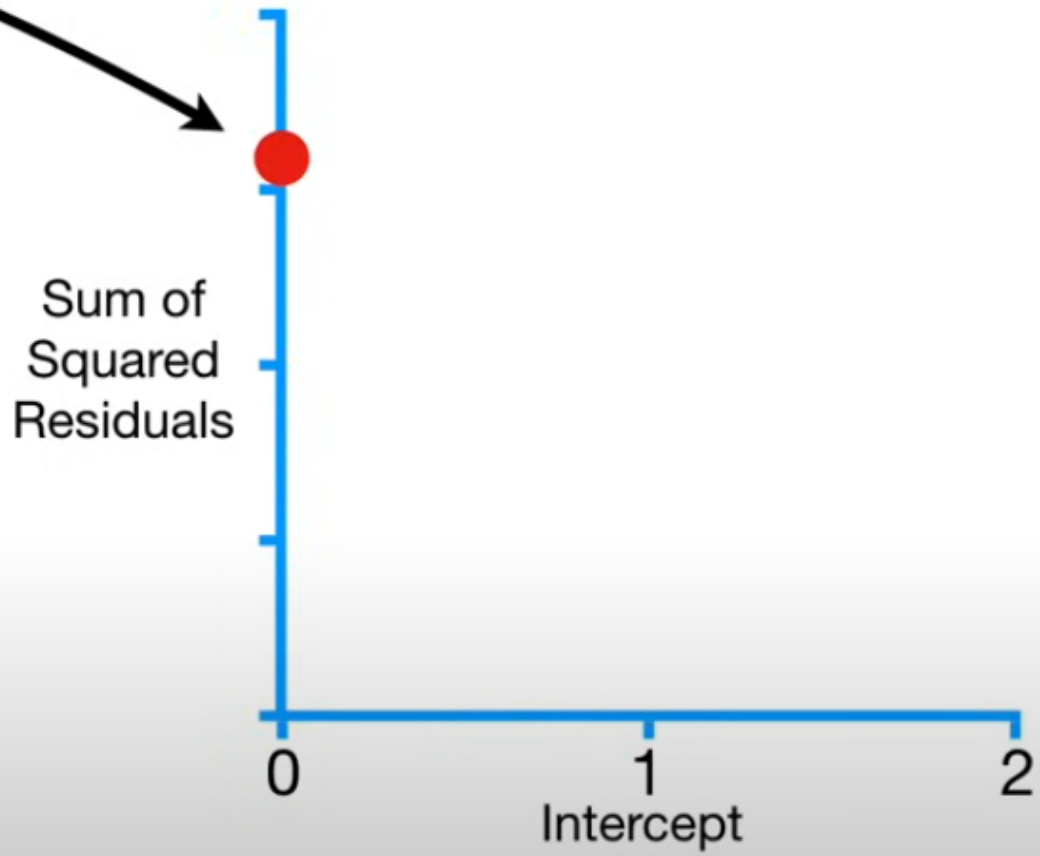
So let's get back to using **Gradient Descent** to find the optimal value for the **Intercept**, starting from a random value. In this case, the random value was **0**.



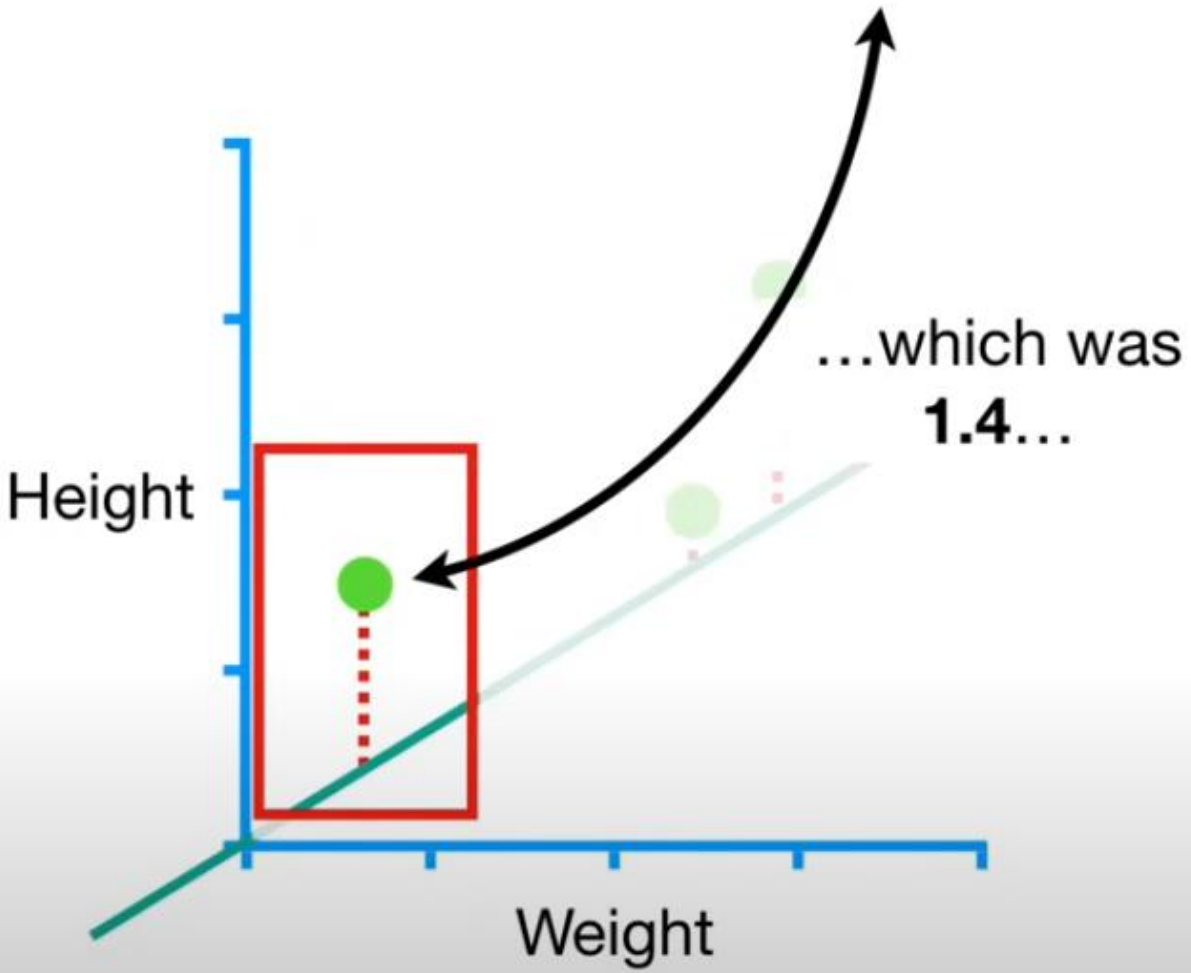
Sum of squared residuals

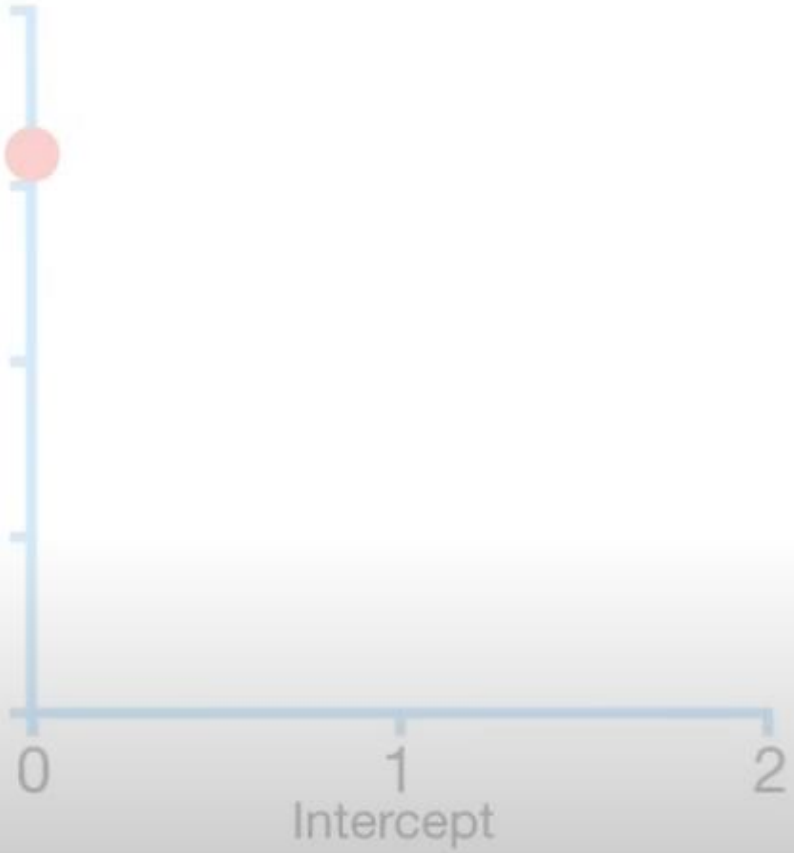
When we calculated the Sum of the Squared Residuals…





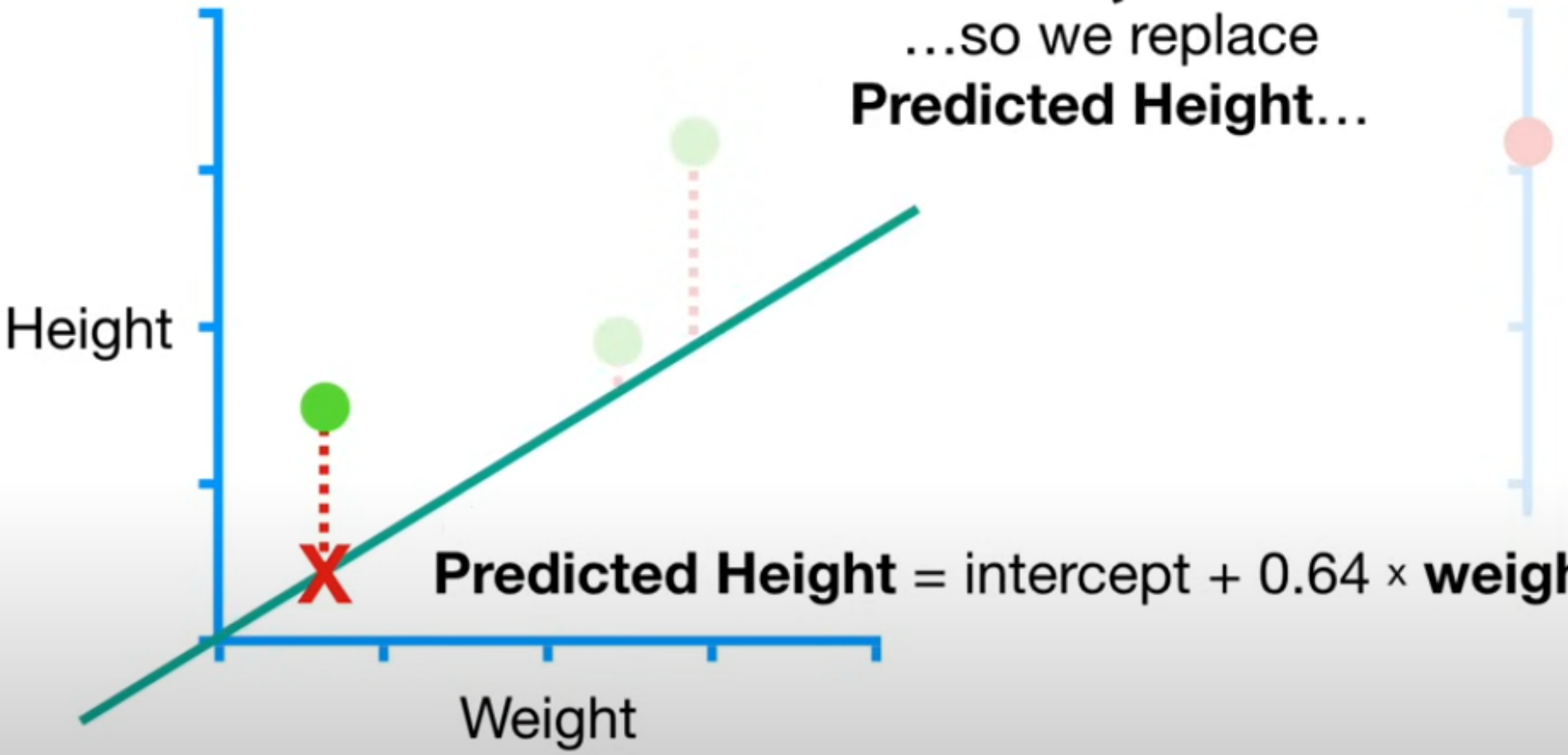
Sum of squared residuals = (**1.4** - predicted)^2





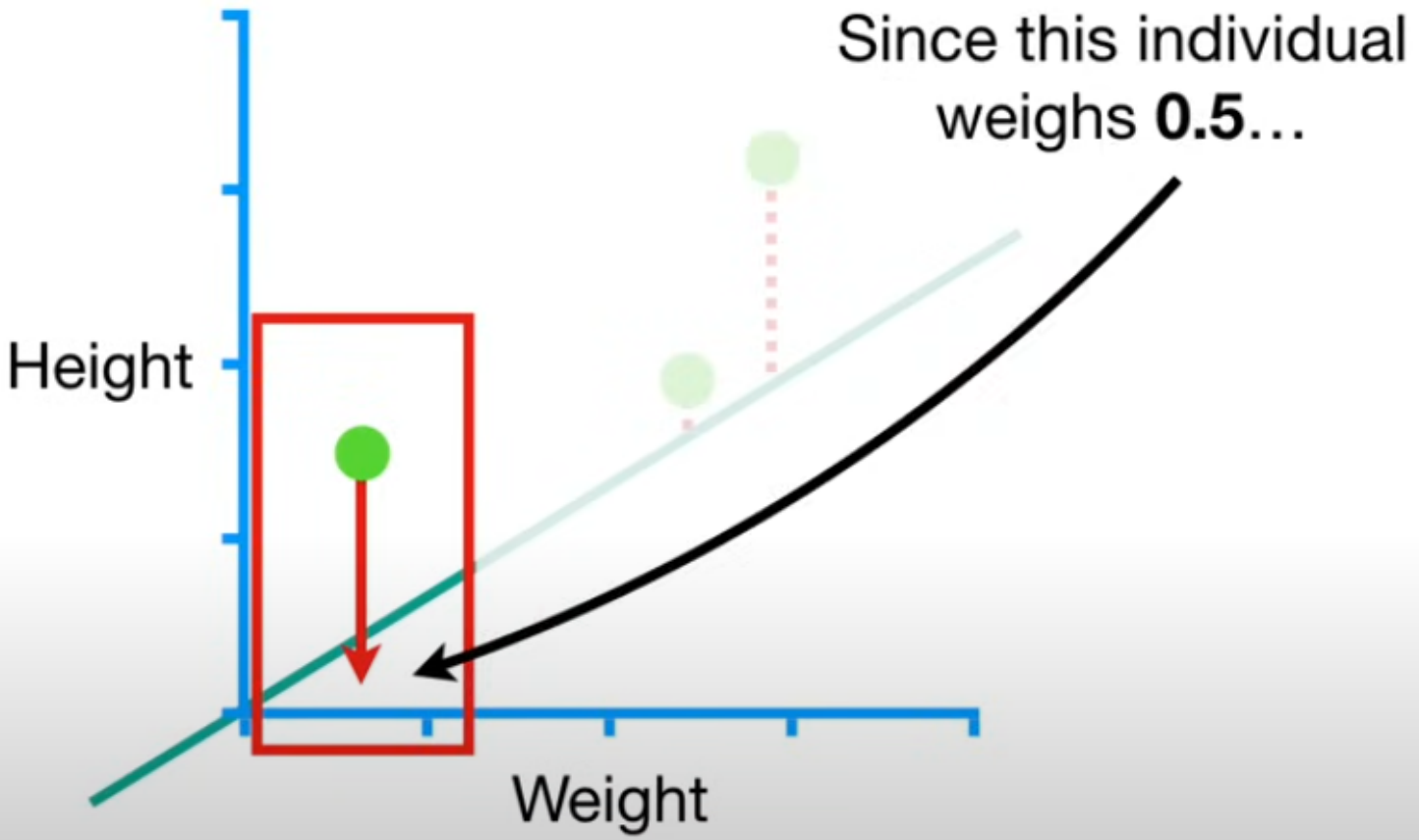
Sum of squared residuals = (**1.4** - predicted)^2

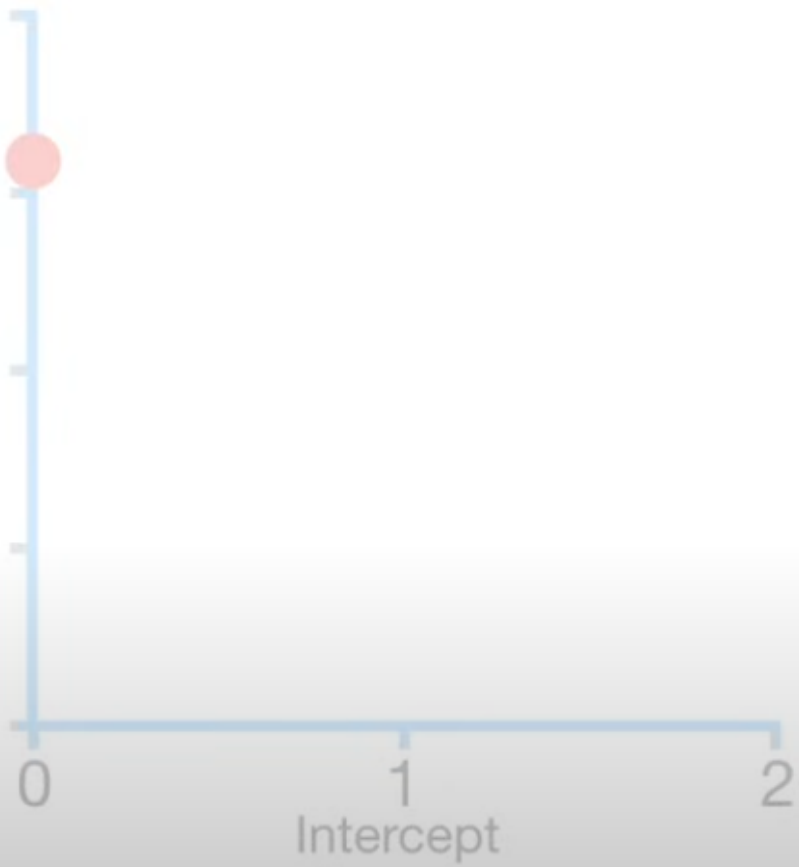
…so we replace
Predicted Height…



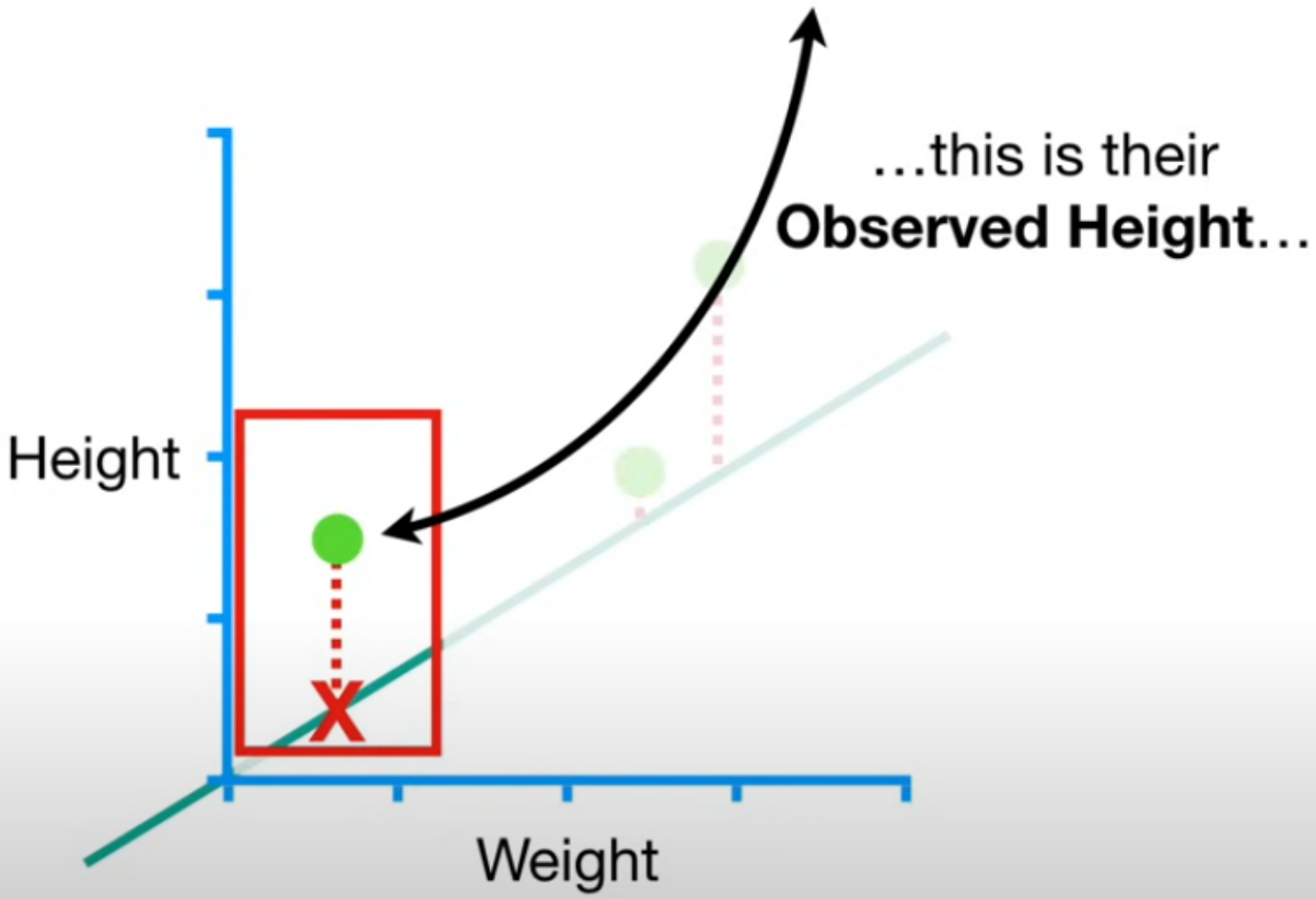
Predicted Height = intercept + 0.64 × **weight**

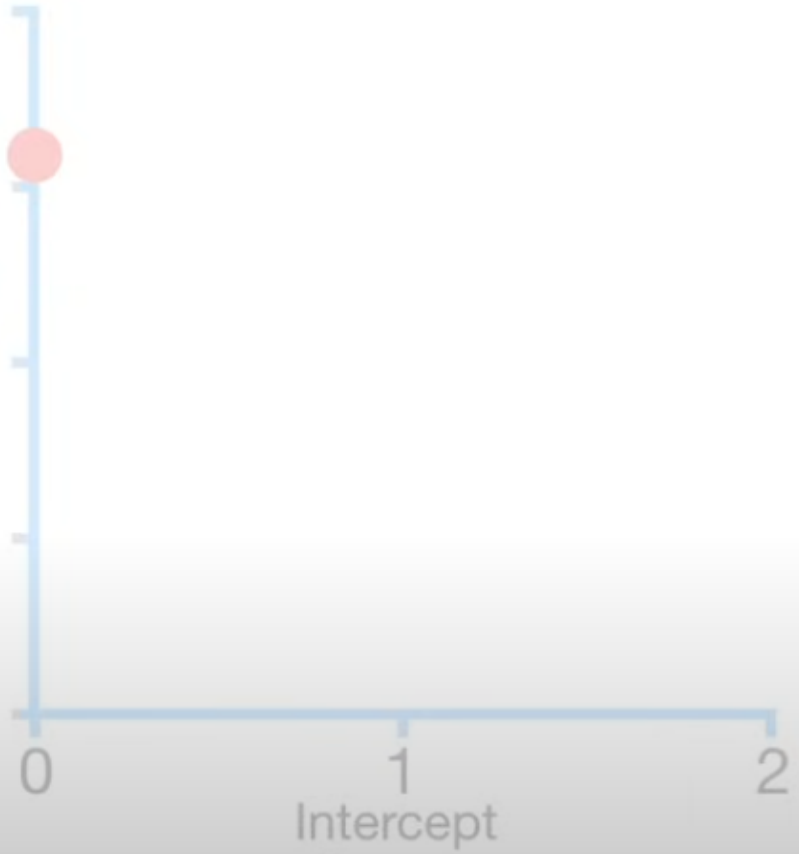
Sum of squared residuals = (**1.4** - (intercept + 0.64 × **weight**))^2





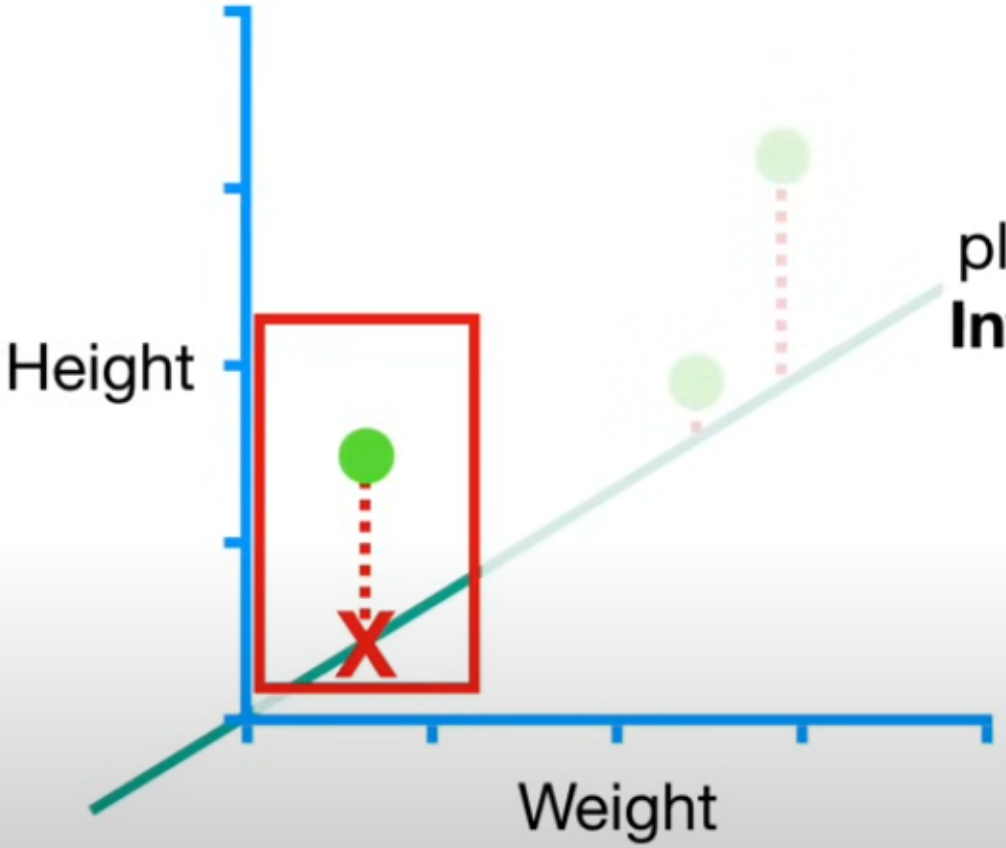
Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

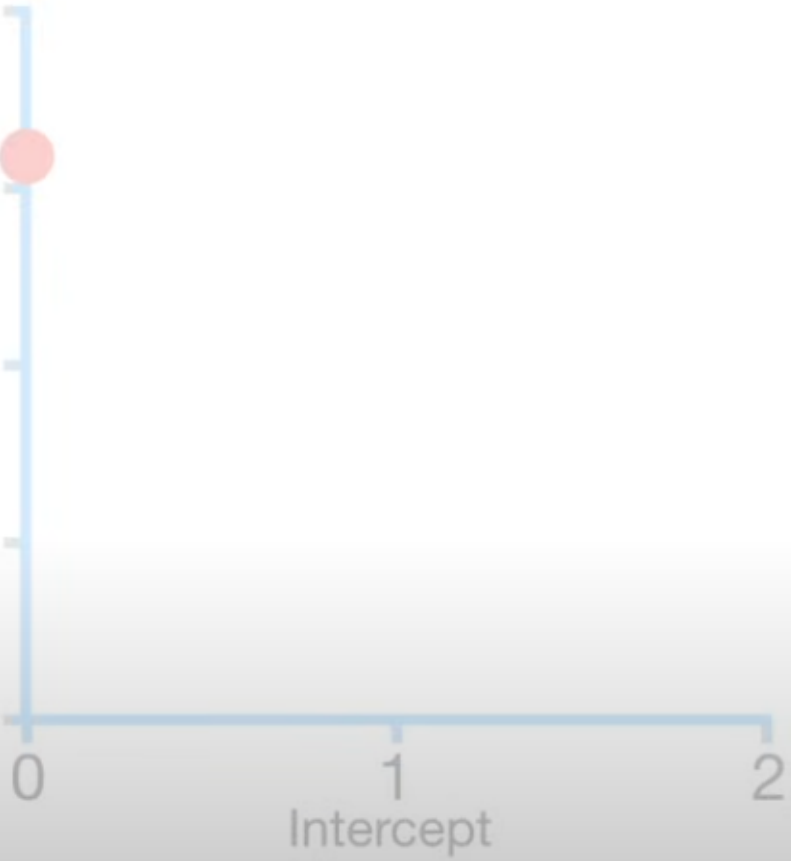




Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

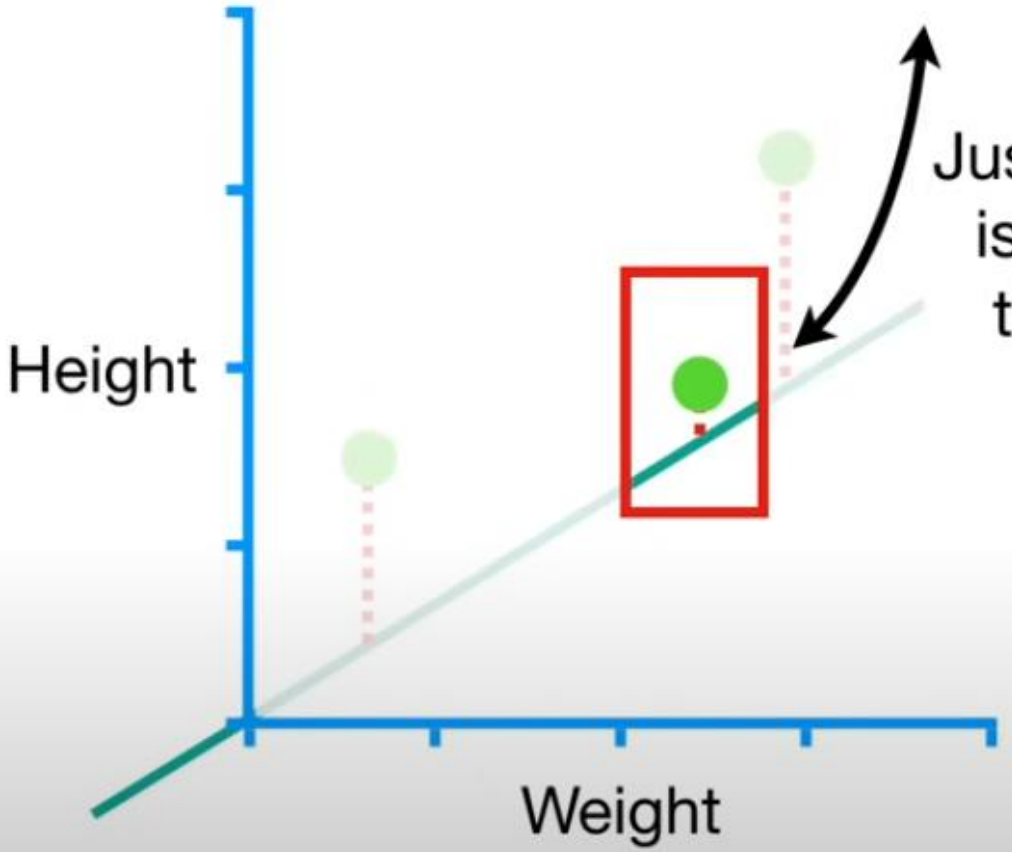
NOTE: We can now plug in any value for the **Intercept** and get a new **Predicted Height**.





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

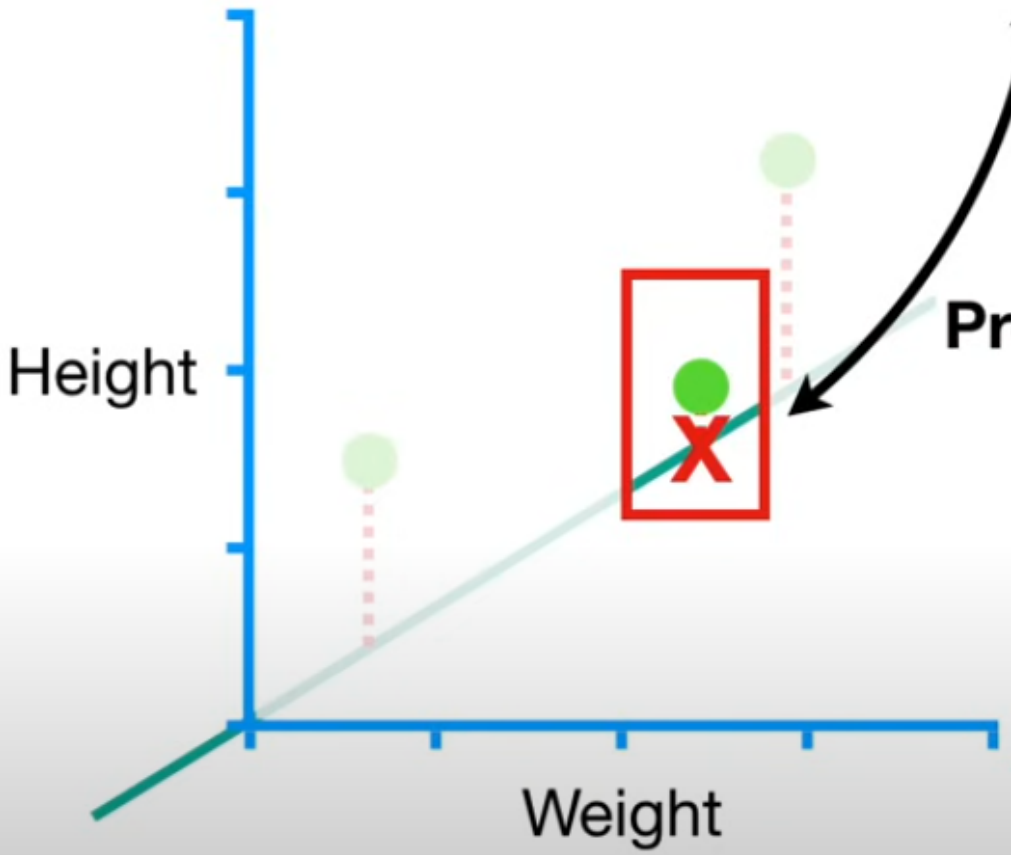
+ (observed - predicted)^2



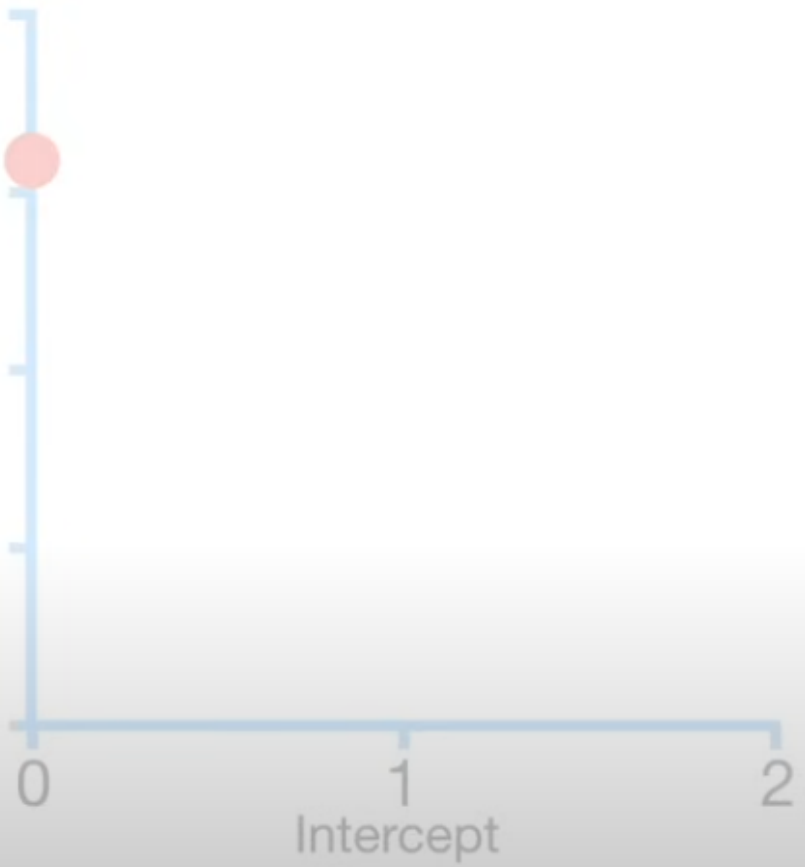
Just like before, the residual is the difference between the **Observed Height**…

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

+ (**1.9** - predicted)^2

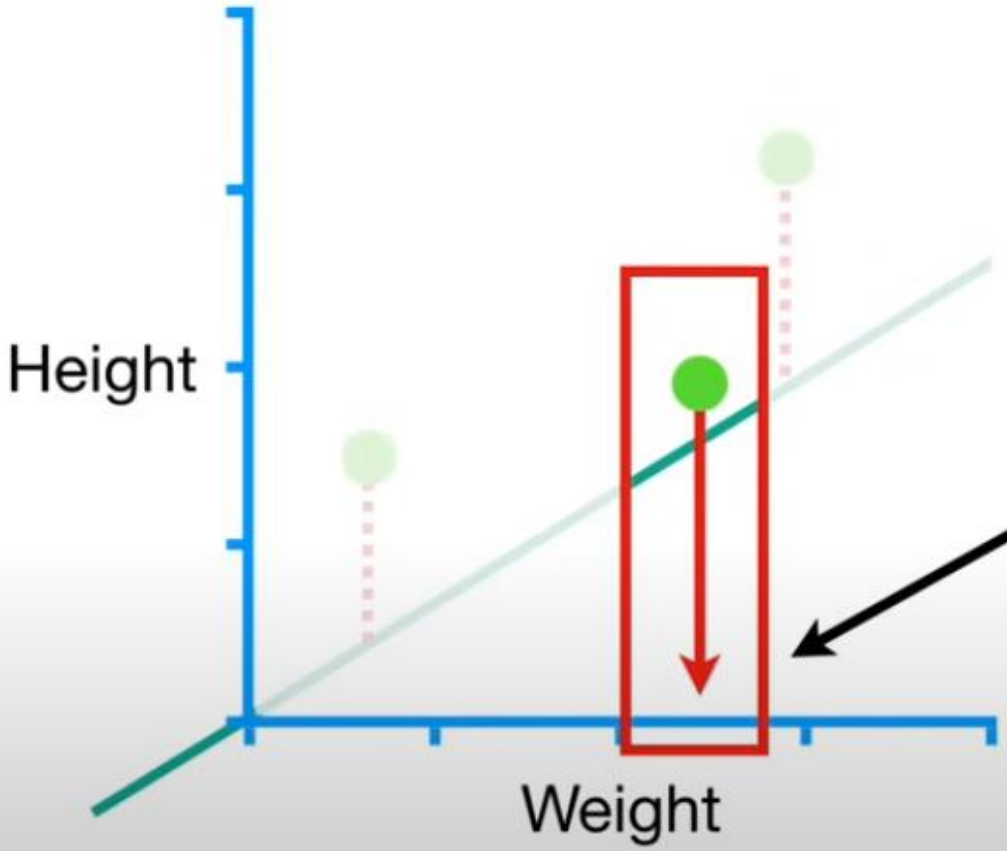


…and the
Predicted Height…

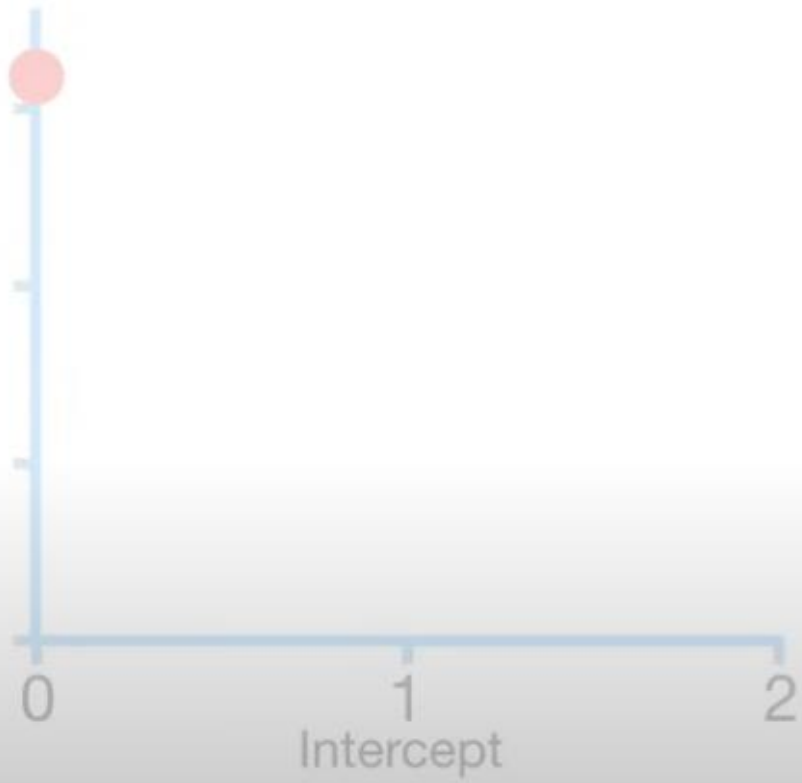


Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

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

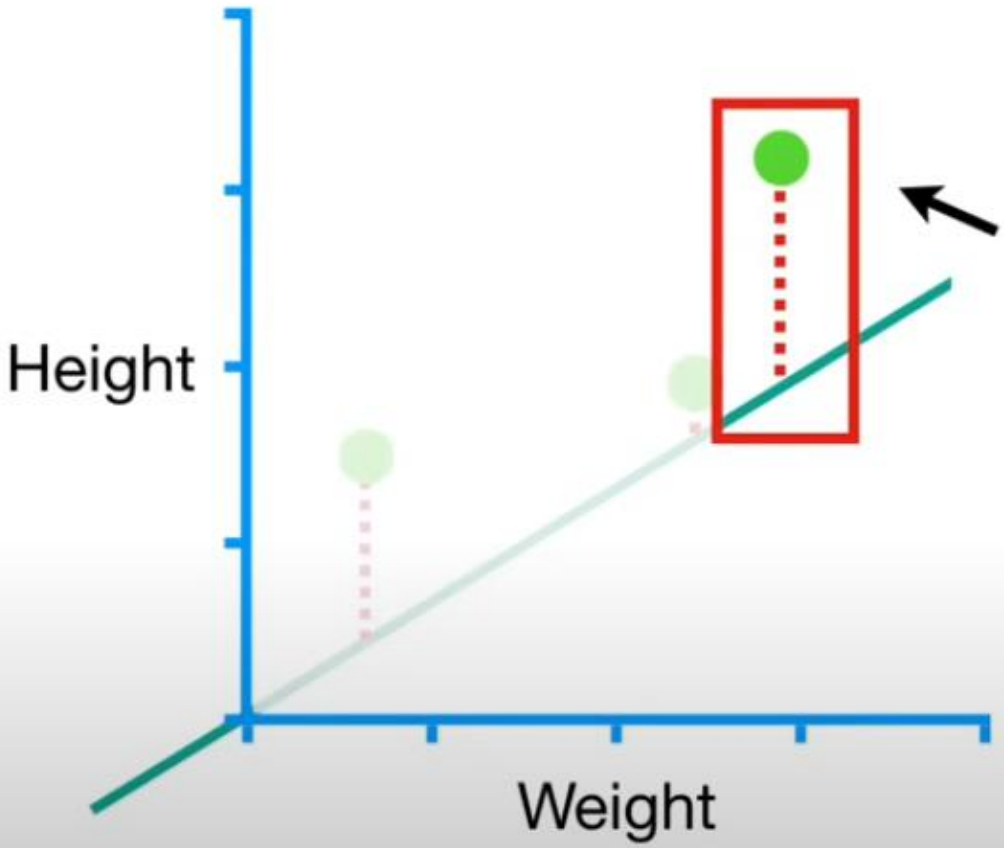


…and since this individual weighs **2.3**…

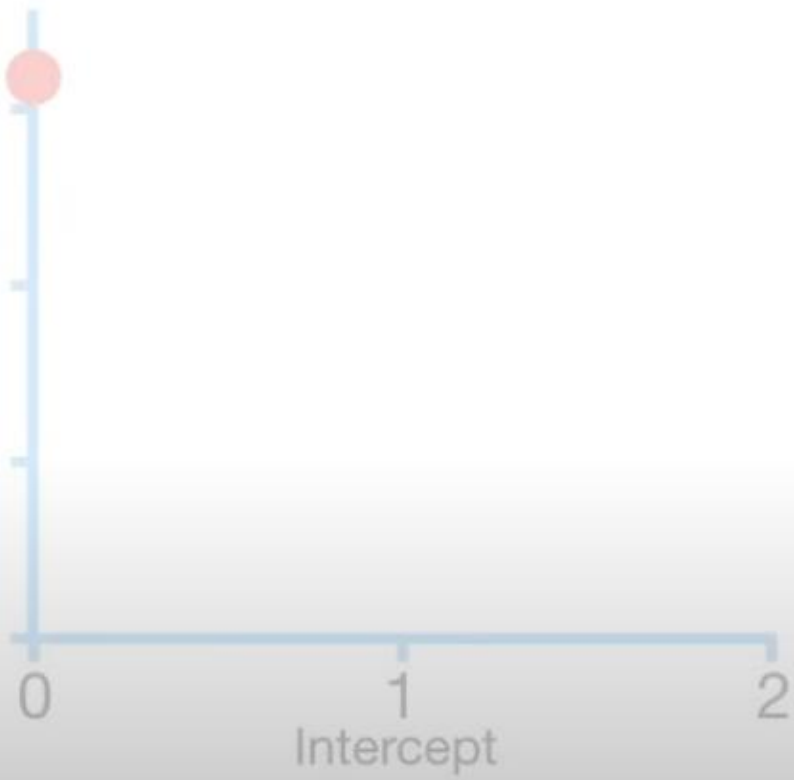


Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

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



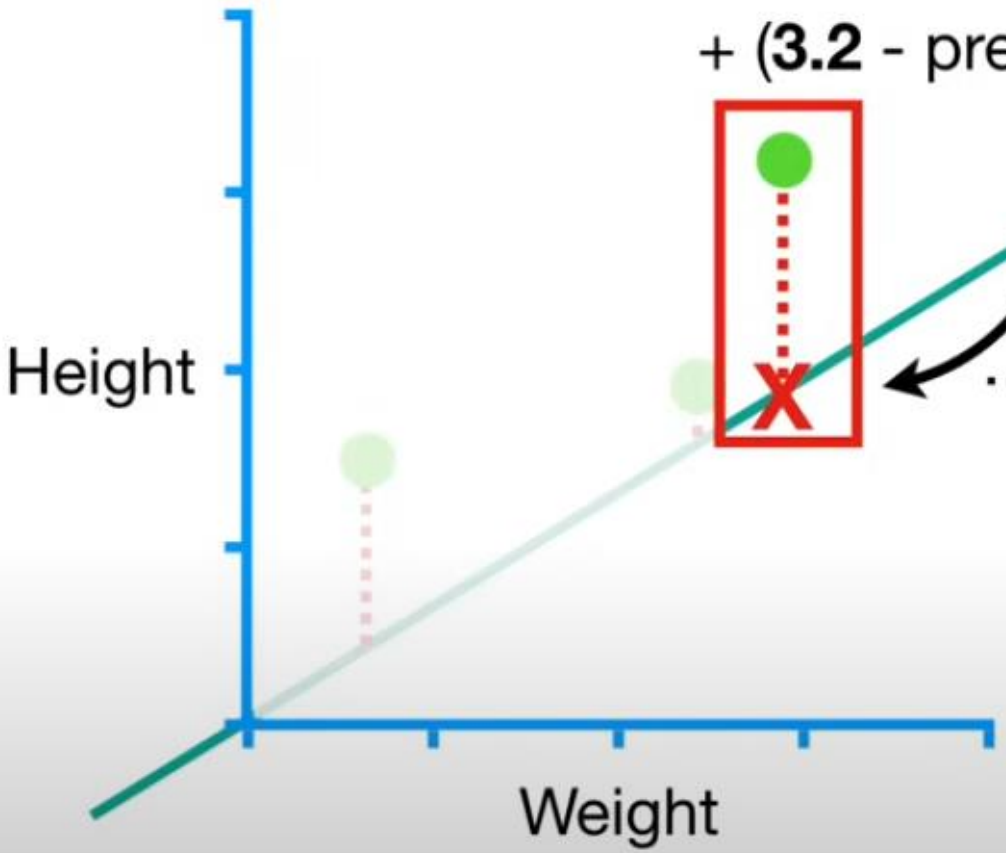
Now let's focus on the last person.



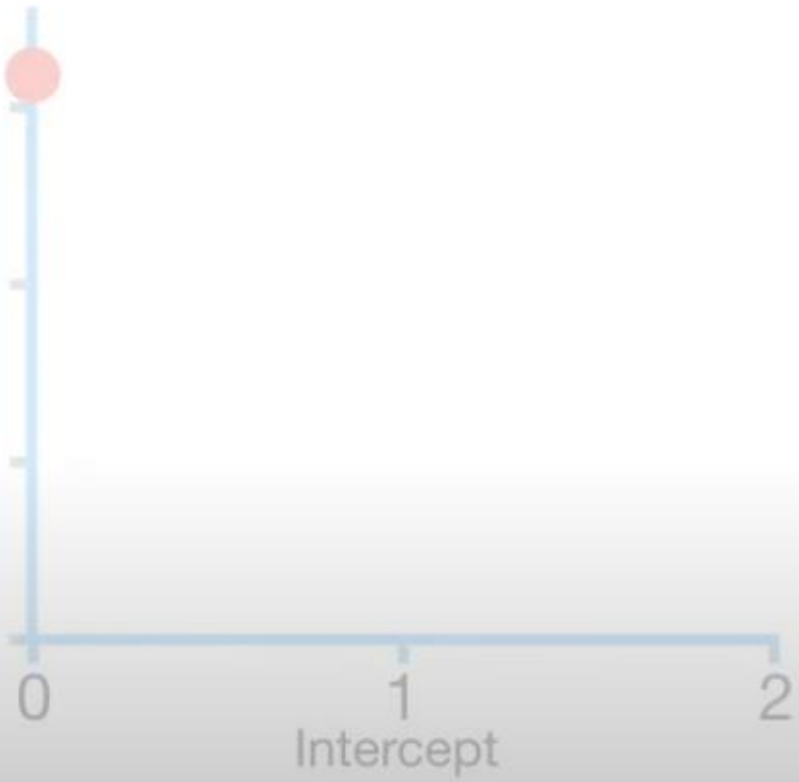
Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

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

+ (**3.2** - predicted)^2



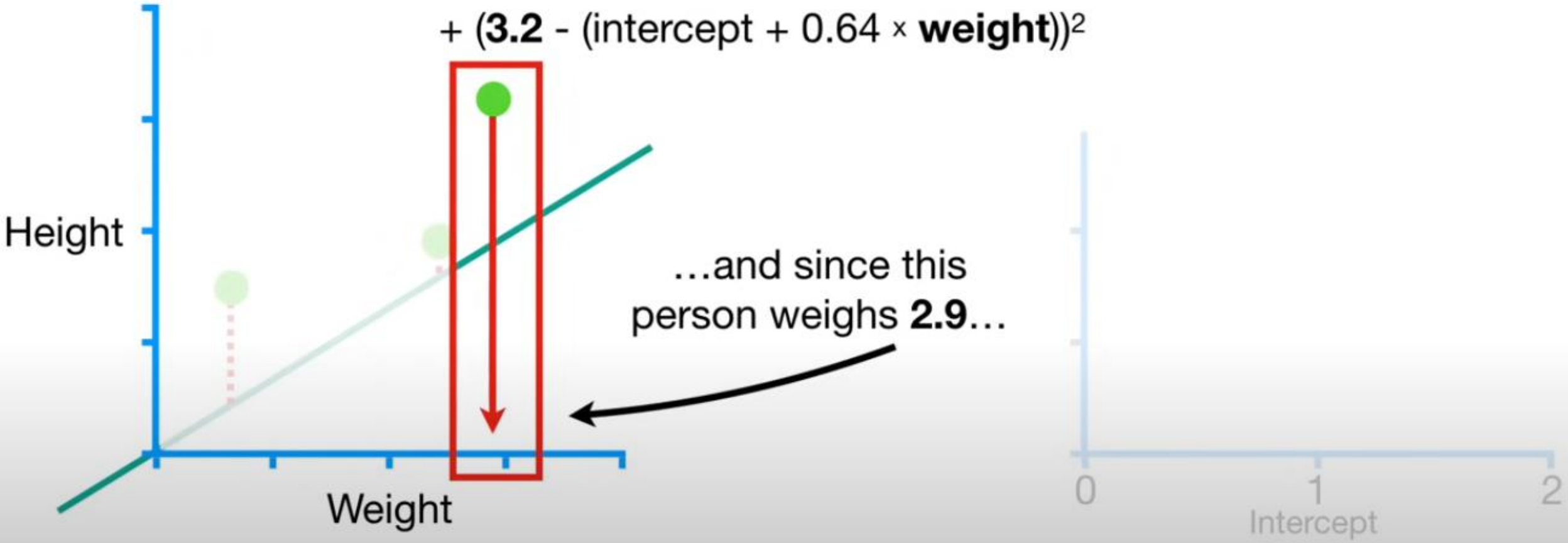
…and the **Predicted Height**…



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

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

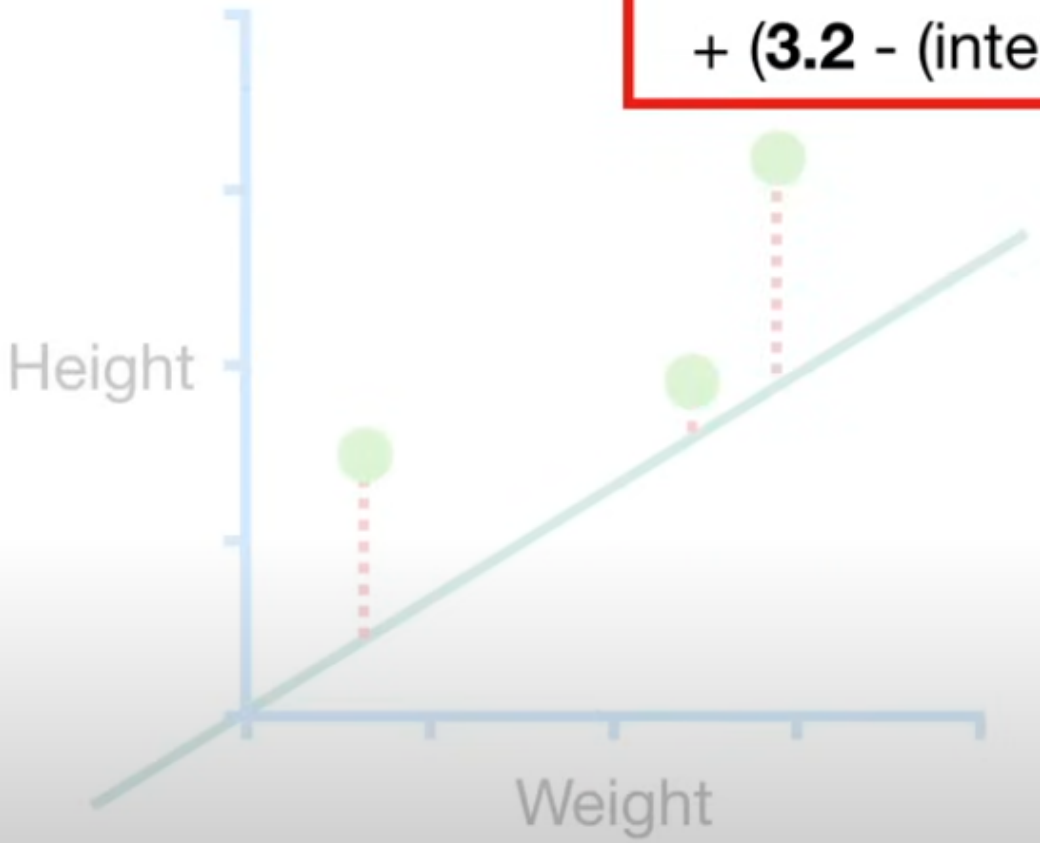
+ (**3.2** - (intercept + 0.64 × **weight**))^2



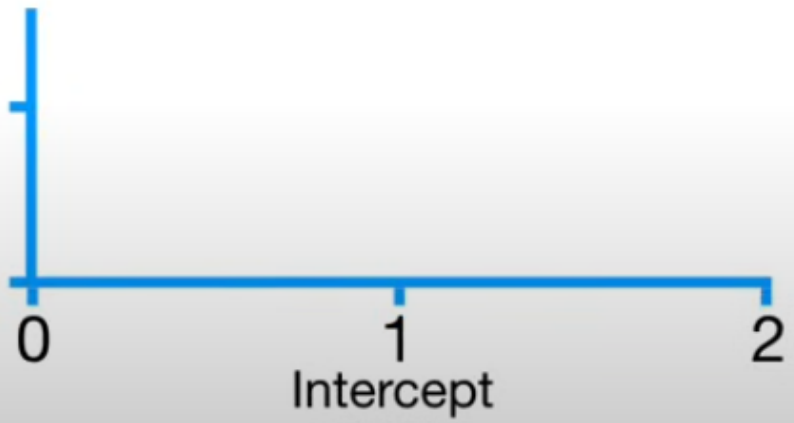
Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

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

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



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

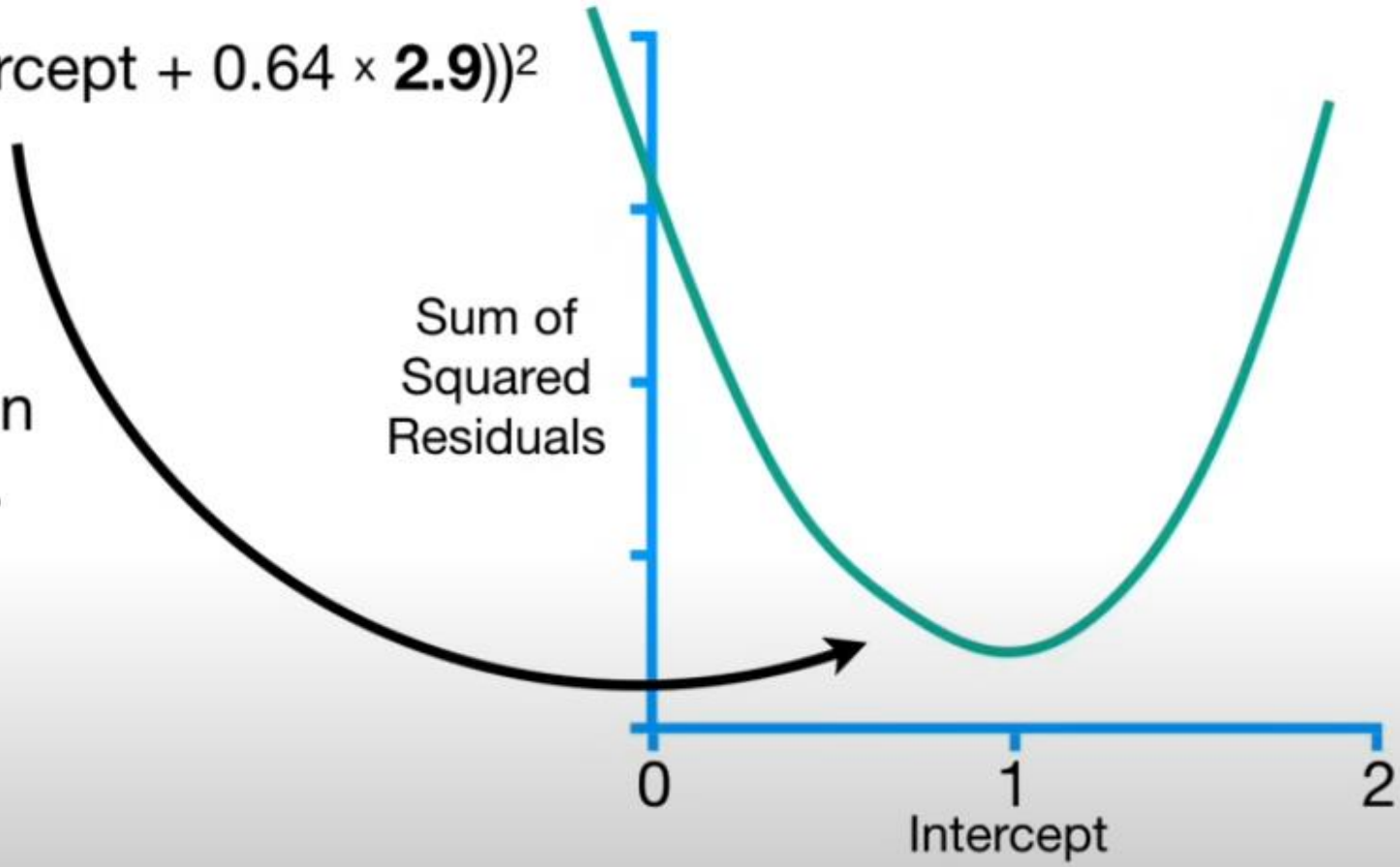


Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

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

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

Thus, we now have an equation for this curve…

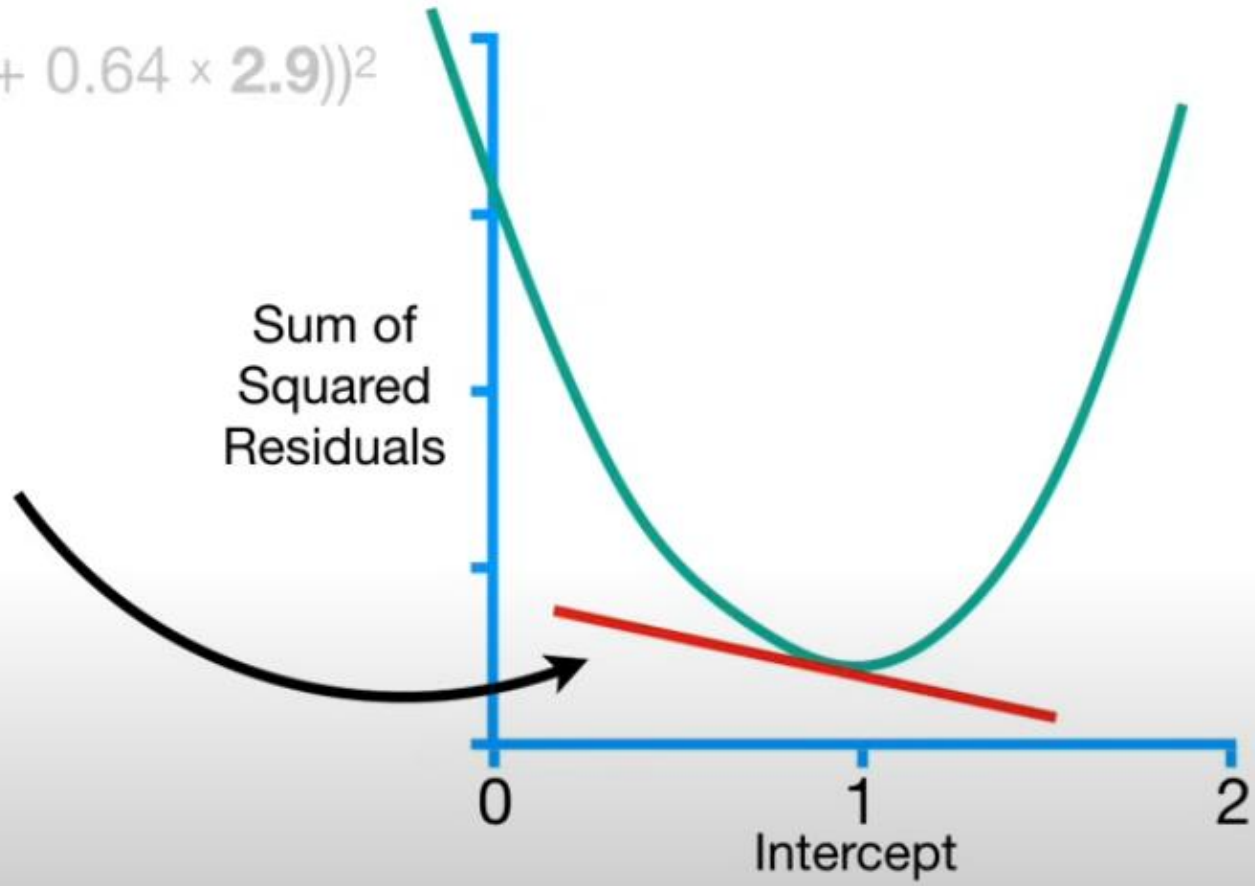


Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

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

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

…and we can take the derivative of this function and determine the slope at any value for the **Intercept**.

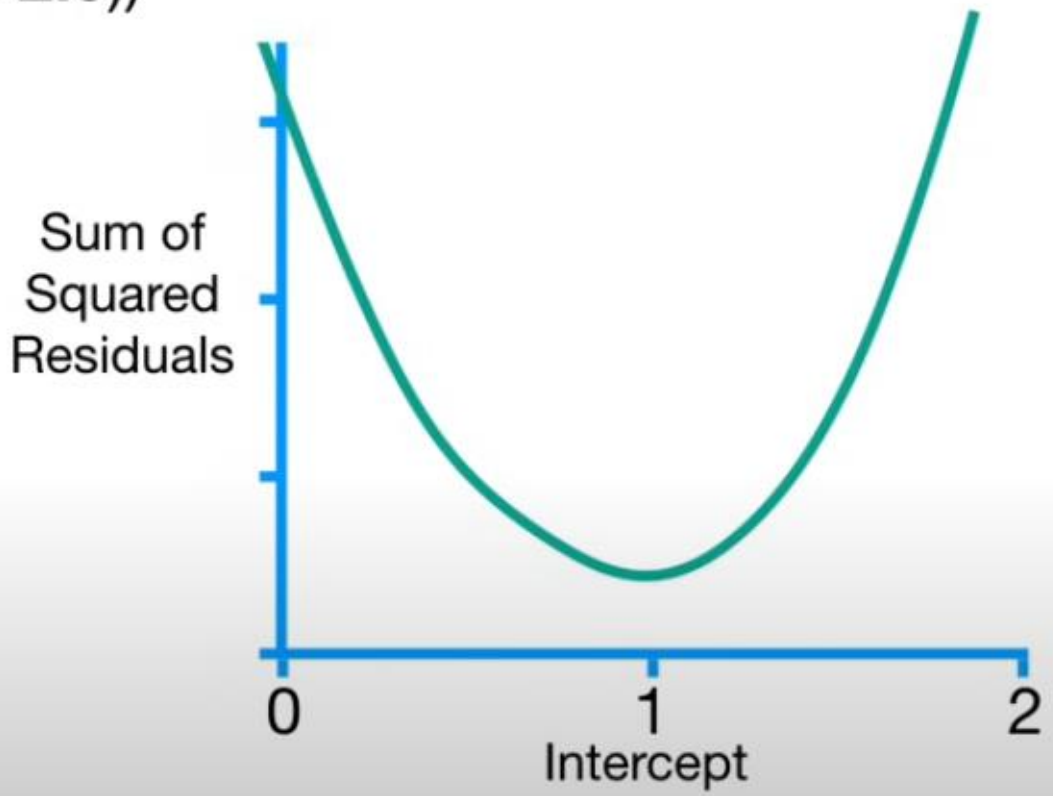


Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

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

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

So let's take the derivative of the Sum of the Squared Residuals with respect to the **Intercept**.



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

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

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

…the derivative of the first part…

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

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

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

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

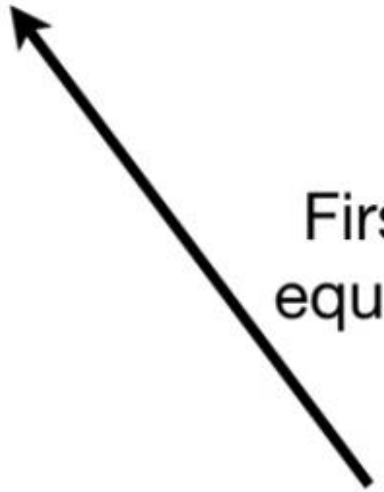
…plus the derivative of the third part.

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

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

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

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



First, we'll move this part of the equation up here so that we have room to work.

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

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

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

The Chain Rule

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

…and multiply that by the derivative of the stuff inside the parentheses.

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

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

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

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

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

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

…and this…

…is the derivative of the first part…

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

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

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

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 =

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

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

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

Let's move the derivative up here so that it's not taking up half of the screen.

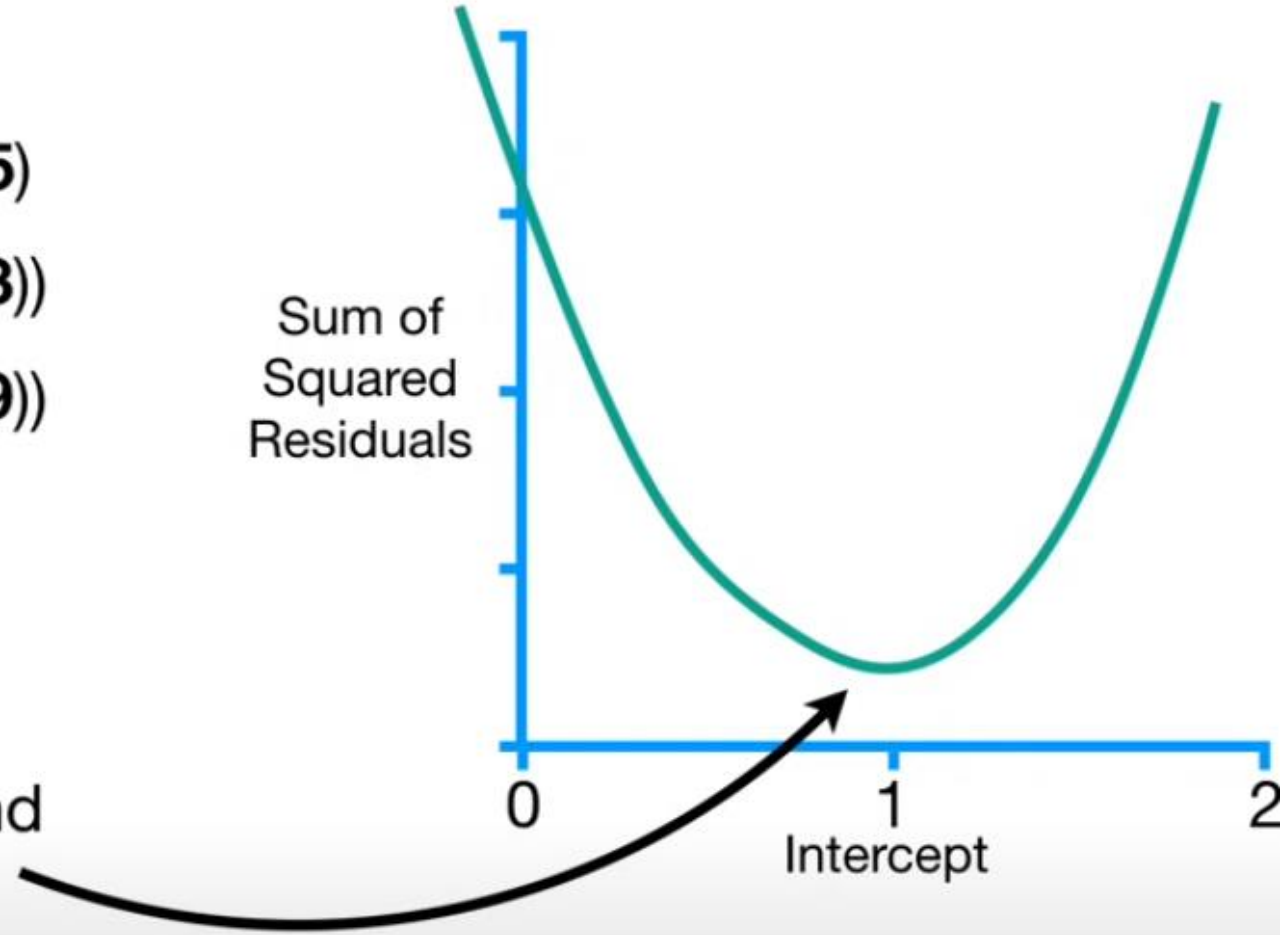
$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

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

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

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

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



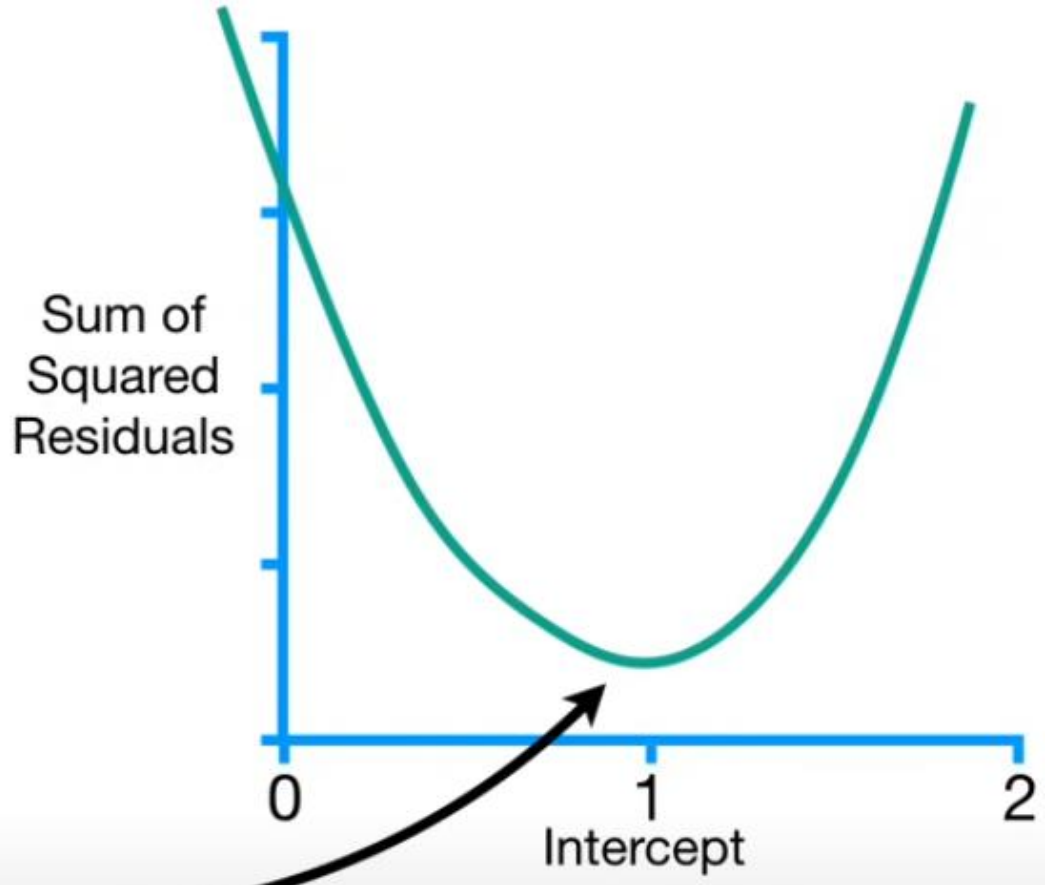
$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

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

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

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

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**.



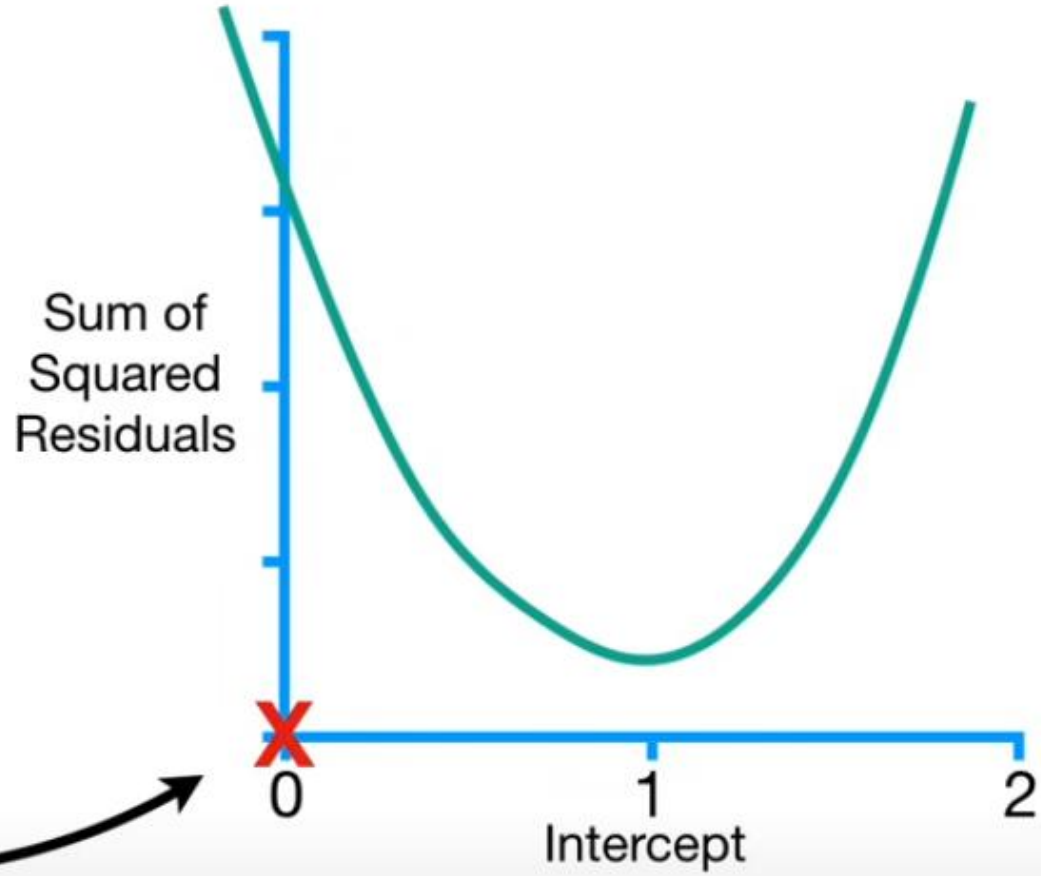
$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

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

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

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

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

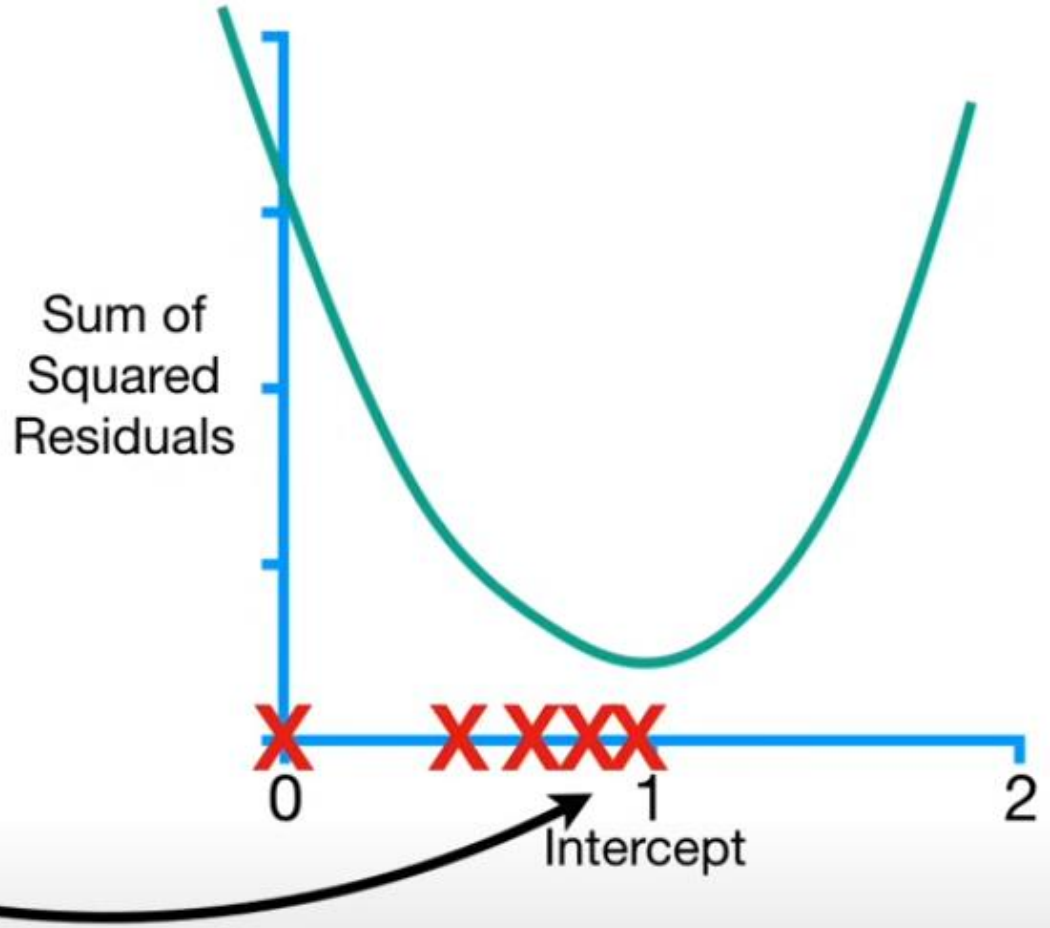


$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

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

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

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



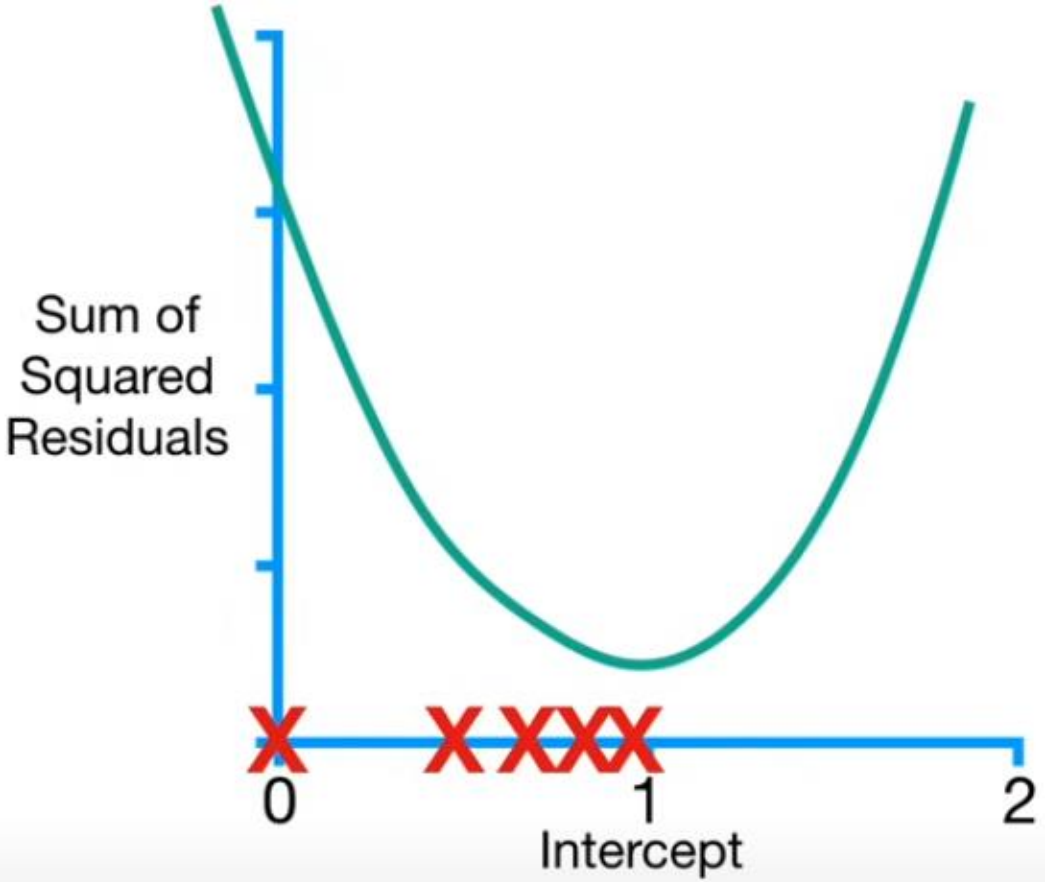
In contrast, **Gradient Descent** finds the minimum value by taking steps from an initial guess until it reaches the best value.

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

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

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

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



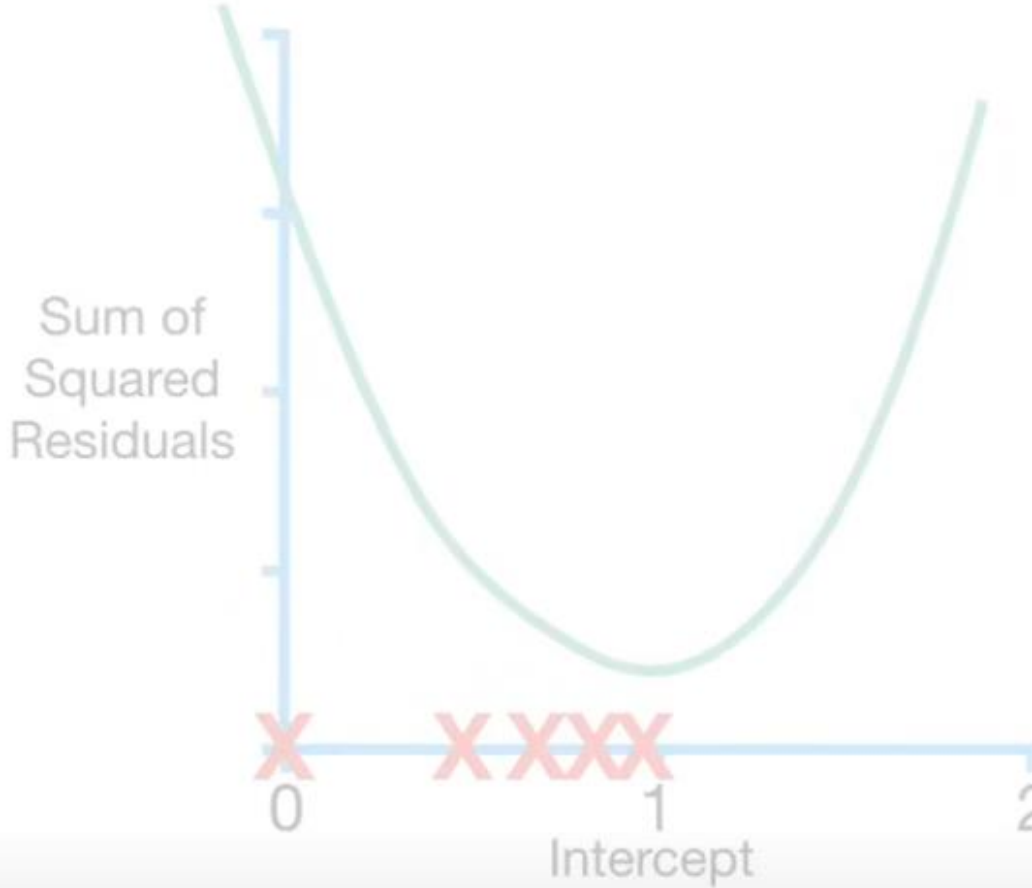
This makes **Gradient Descent** very useful when it is not possible to solve for where the derivative = **0**, and this is why **Gradient Descent** can be used in so many different situations.

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

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

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

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



BAM

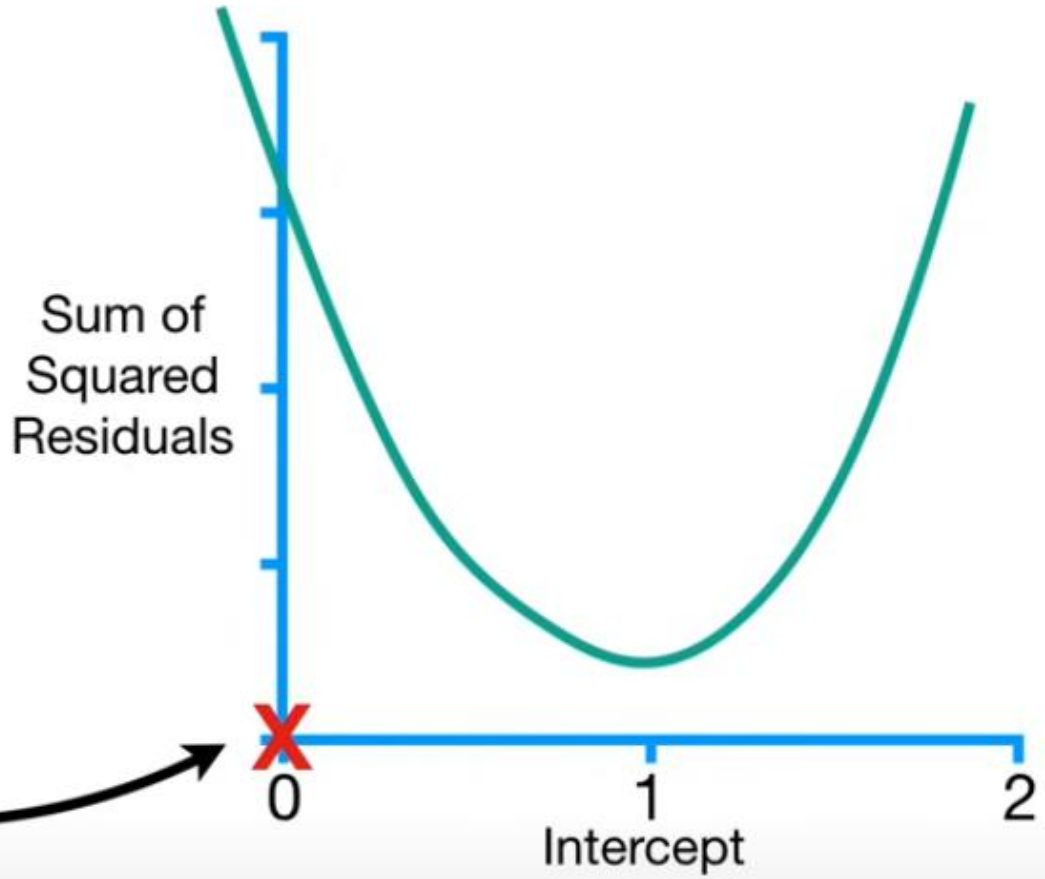
$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

-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**.



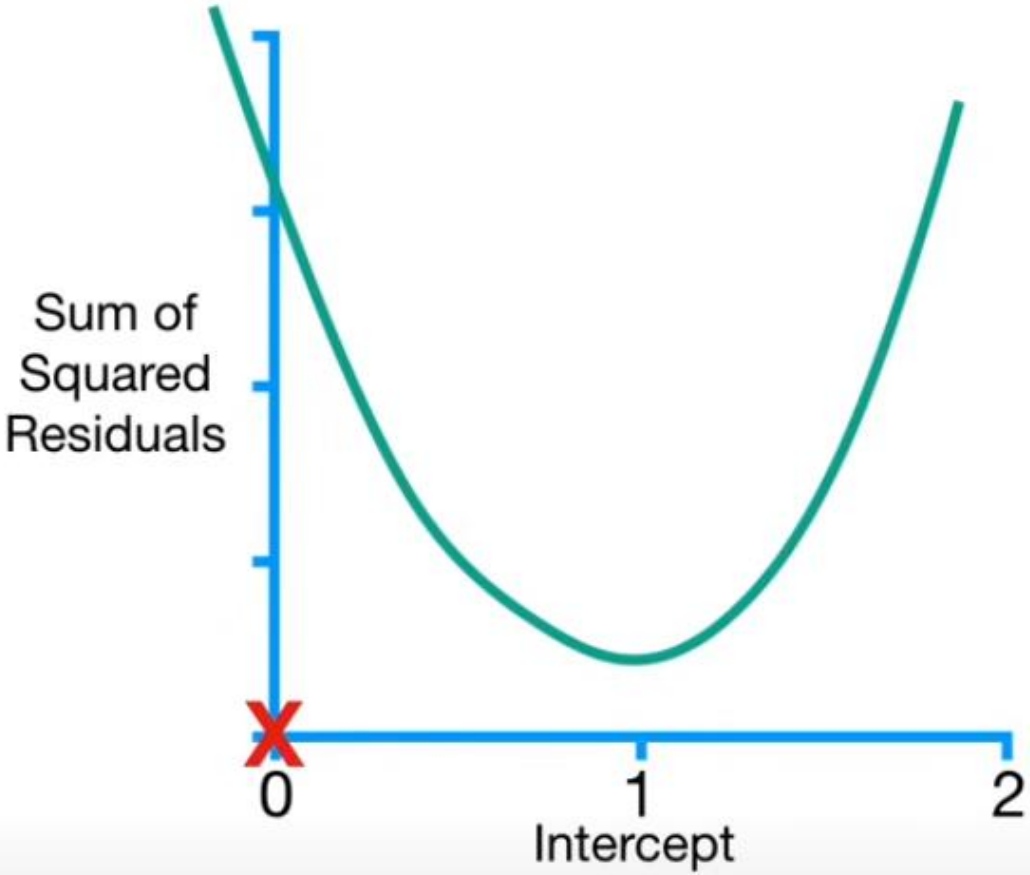
$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

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

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

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

So we plug **0** into the derivative…



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

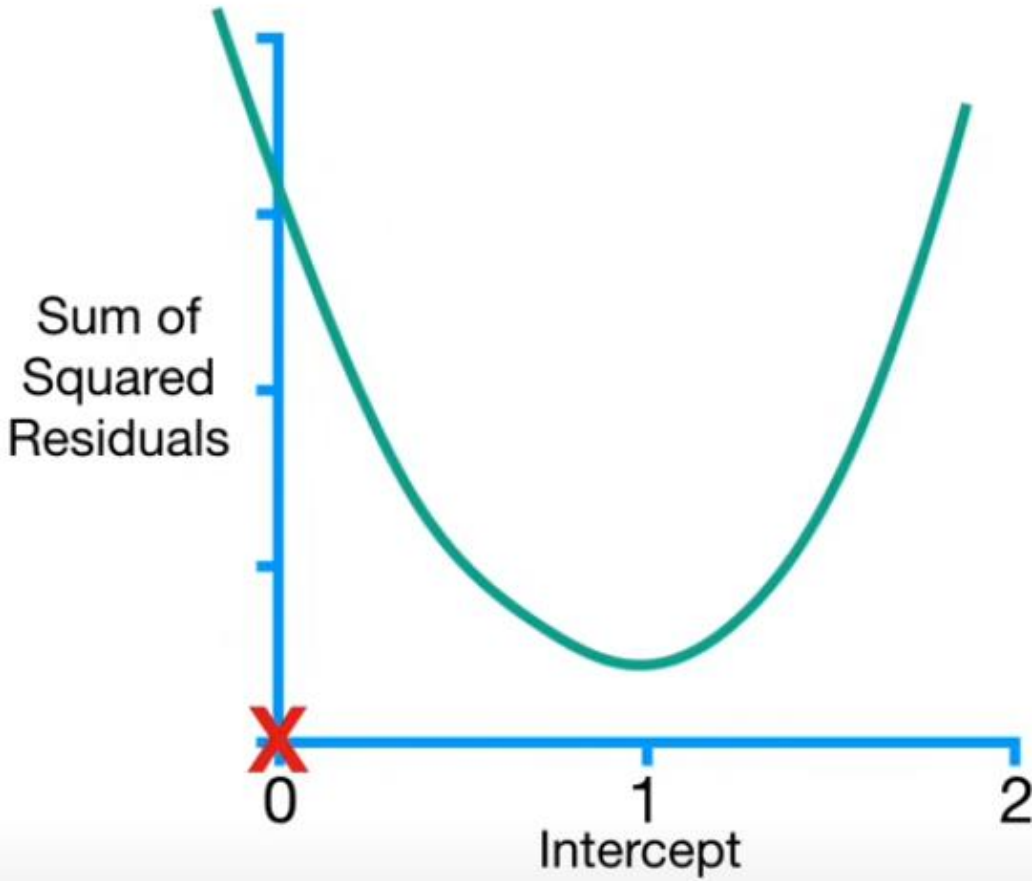
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

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

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

$$= -5.7$$

…and we get **-5.7**.



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

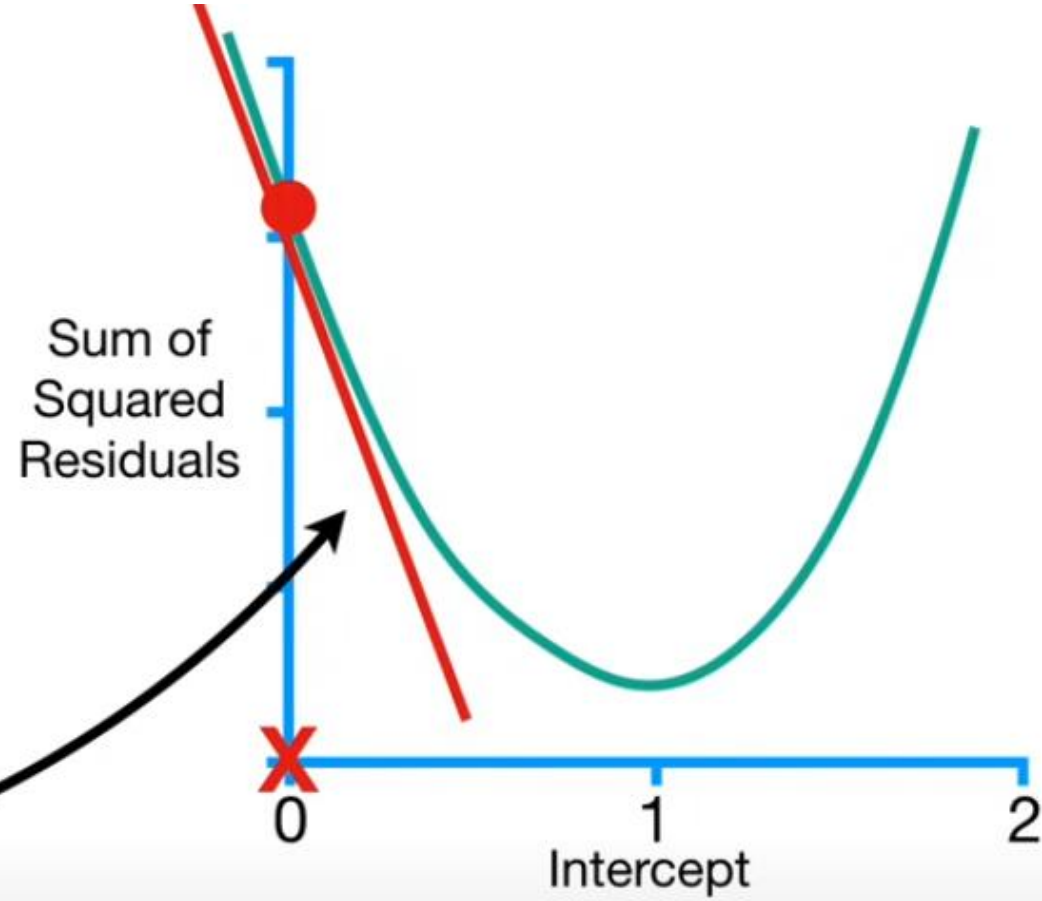
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

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

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

$$= -5.7$$

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



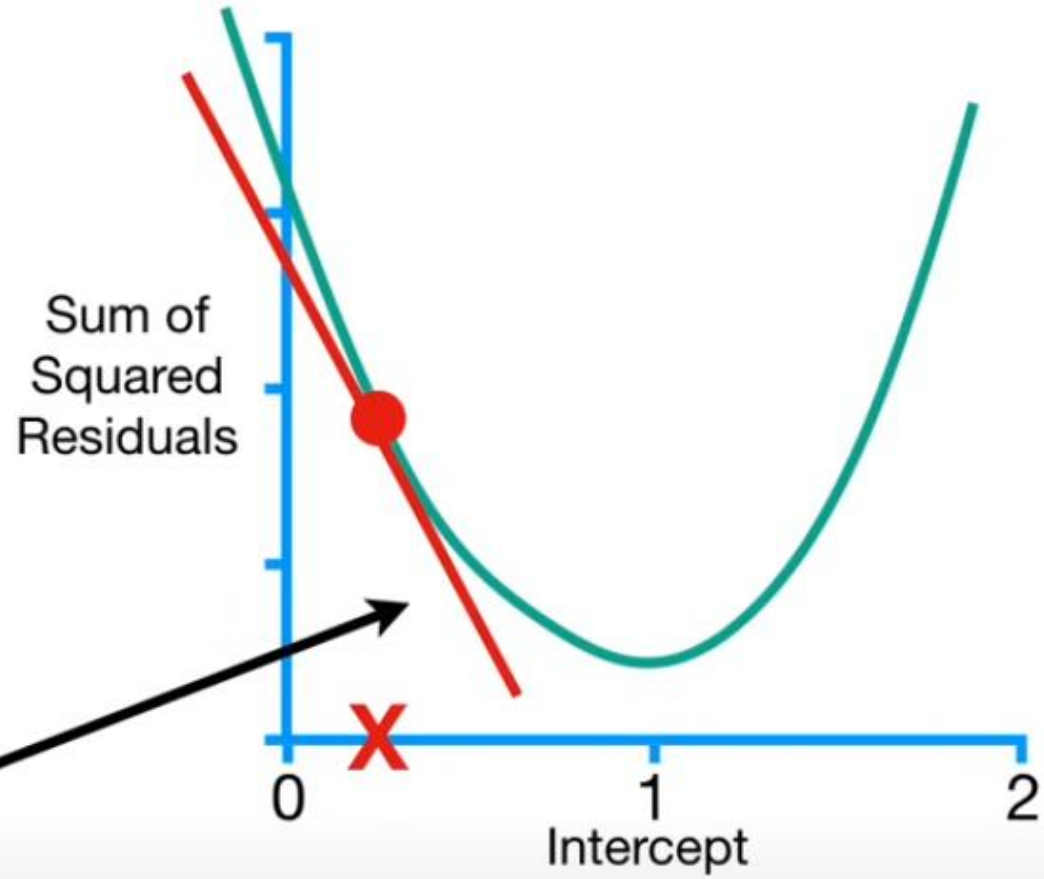
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

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

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

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

$$= -5.7$$



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

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

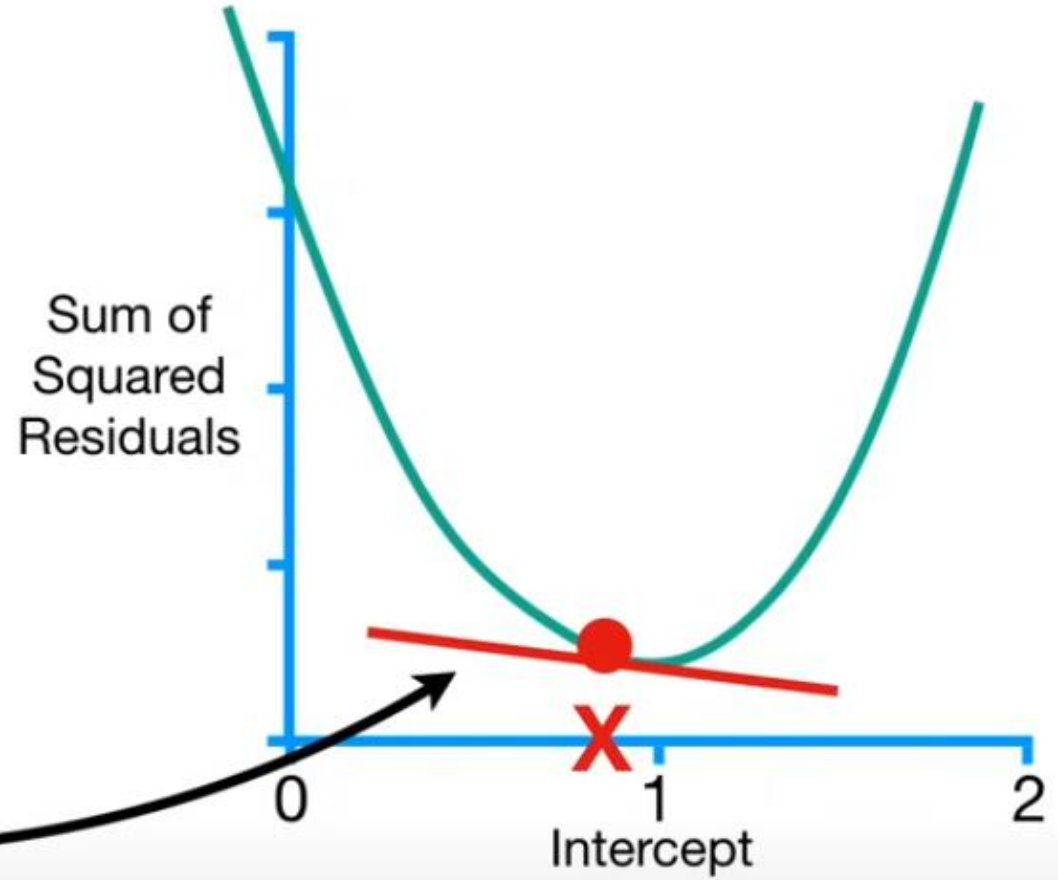
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

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

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

$$= -5.7$$

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



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

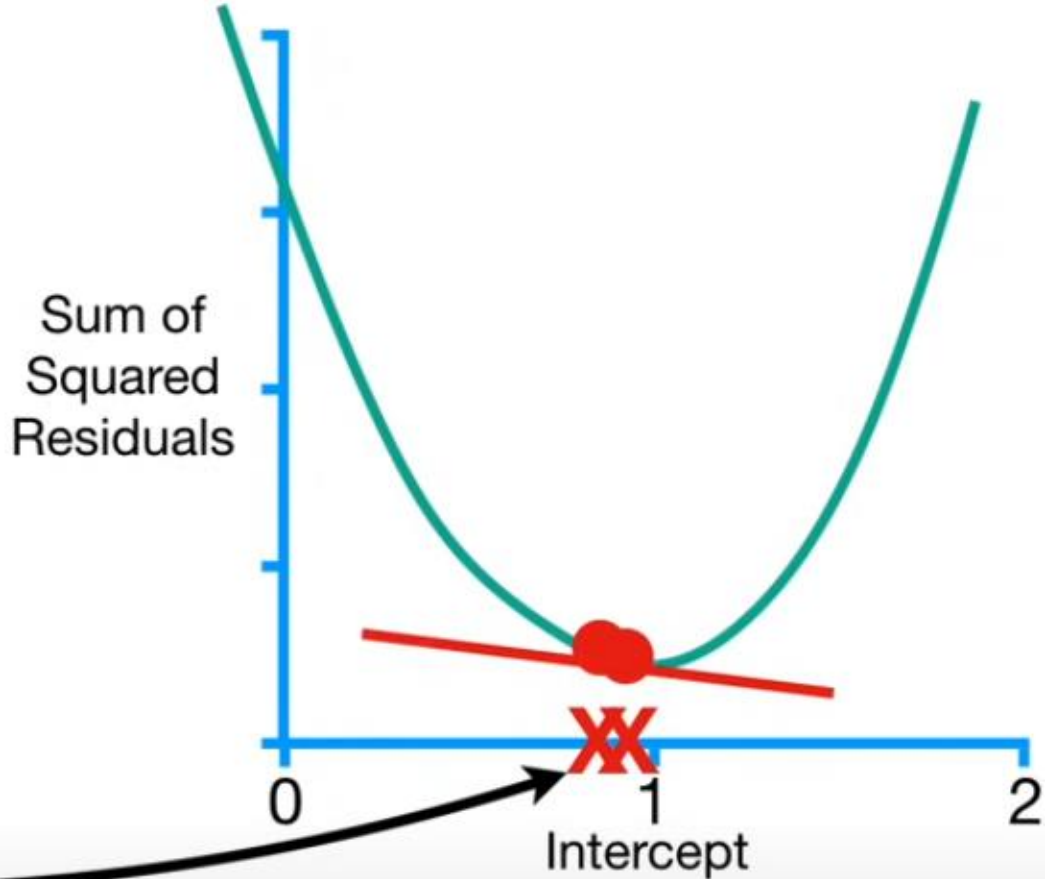
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

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

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

$$= -5.7$$

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



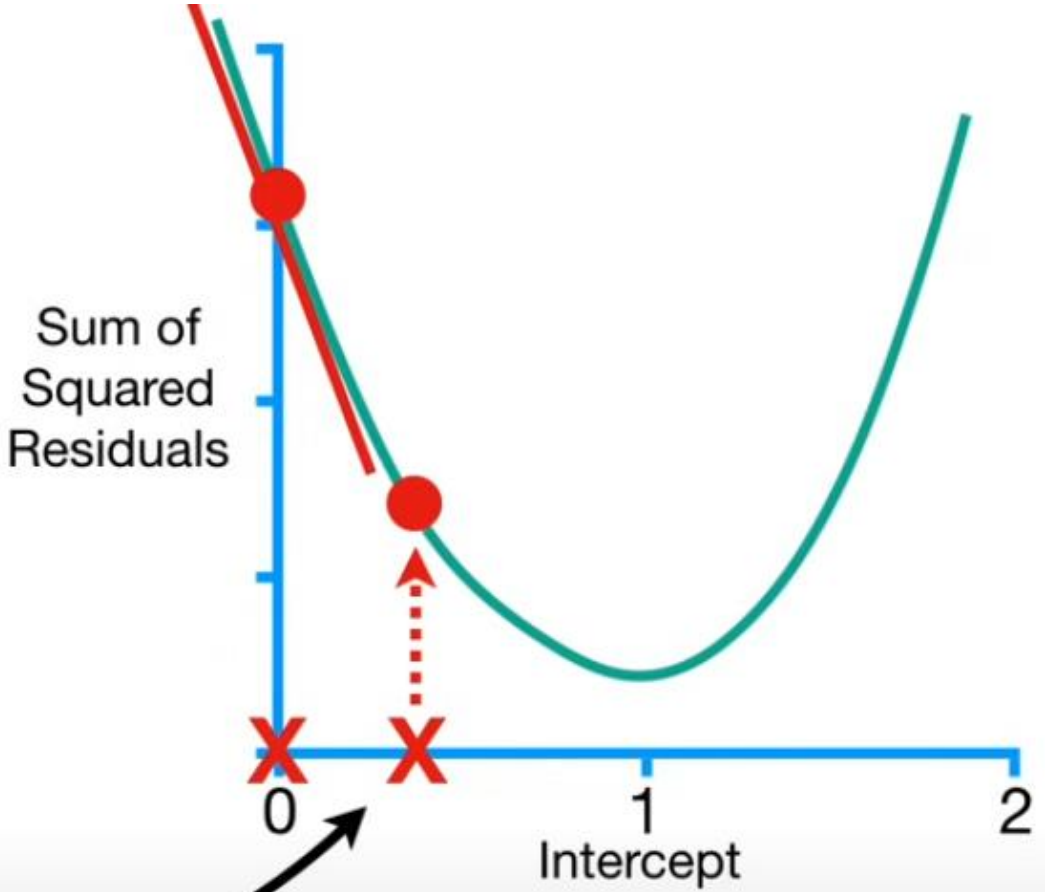
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

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

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

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

$$= -5.7$$



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

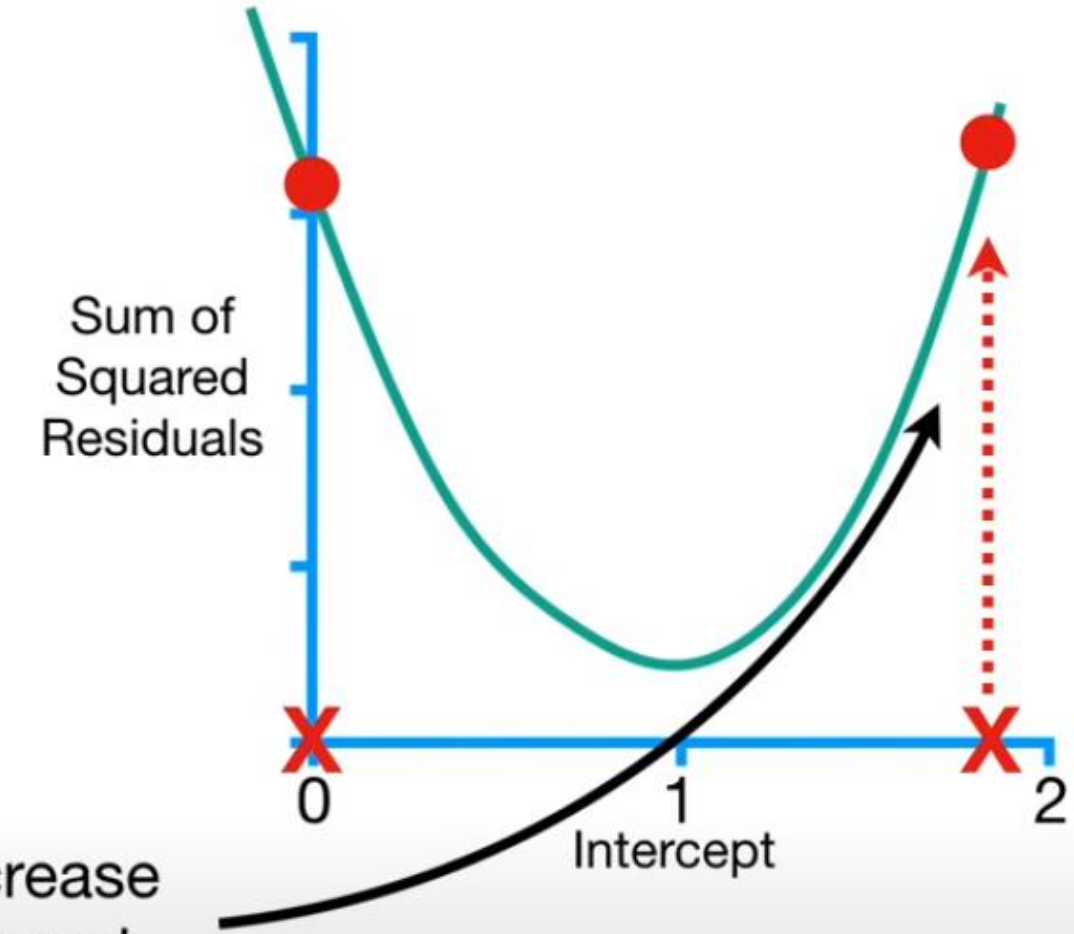
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

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

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

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

$$= -5.7$$



…then we would increase the Sum of the Squared Residuals!

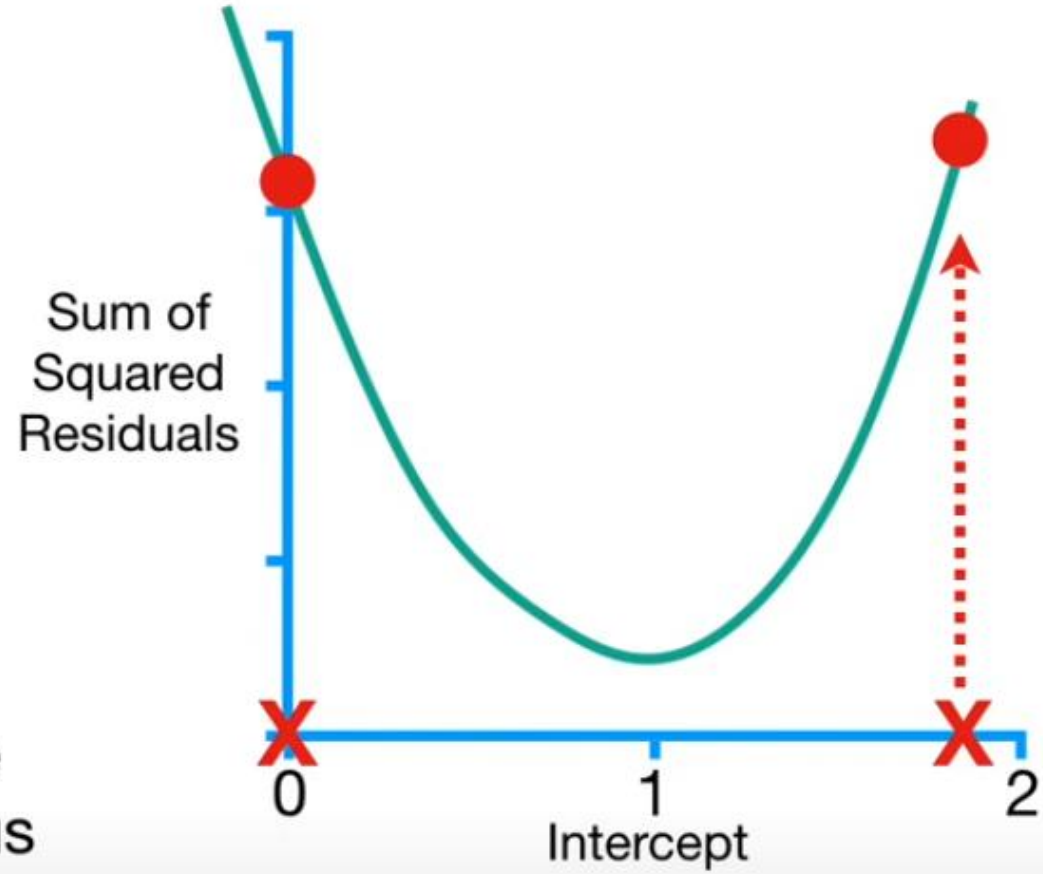
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

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

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

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

$$= -5.7$$



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.

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

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

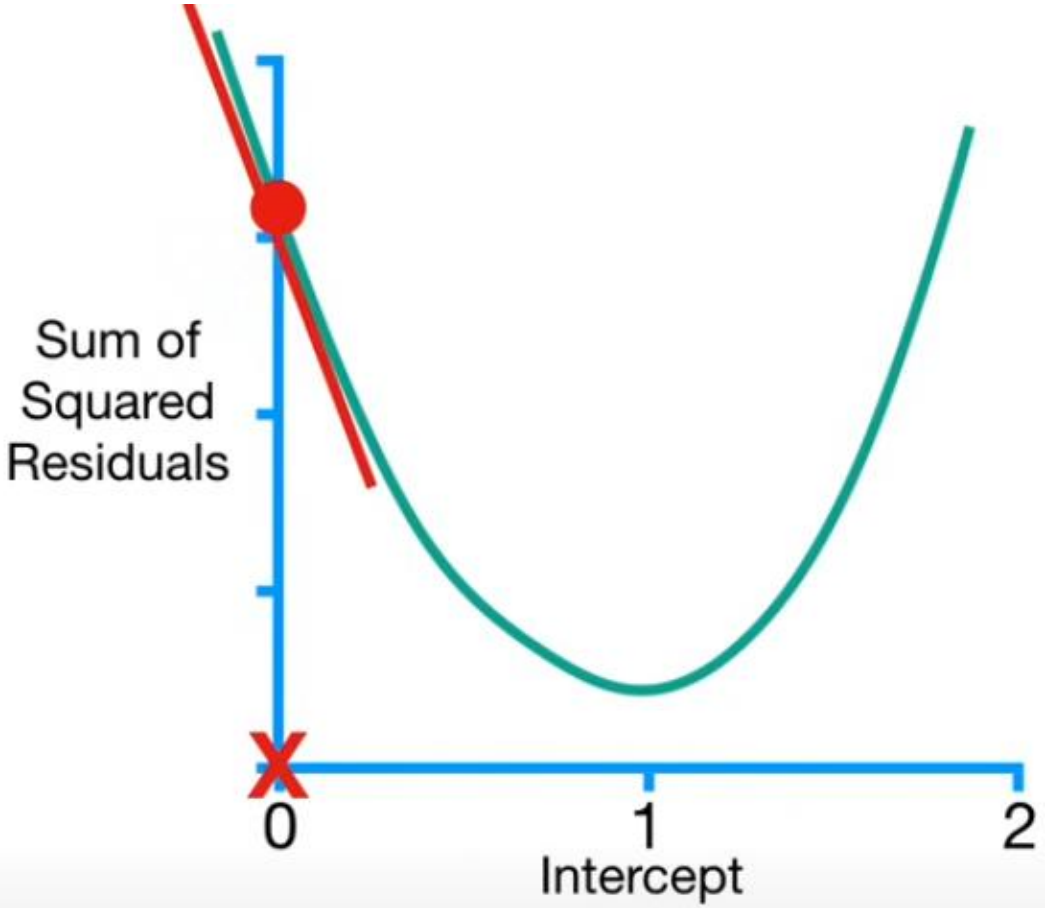
$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

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

$$= -5.7$$

Step Size = -5.7

Gradient Descent determines the **Step Size** by multiplying the **slope**…



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

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

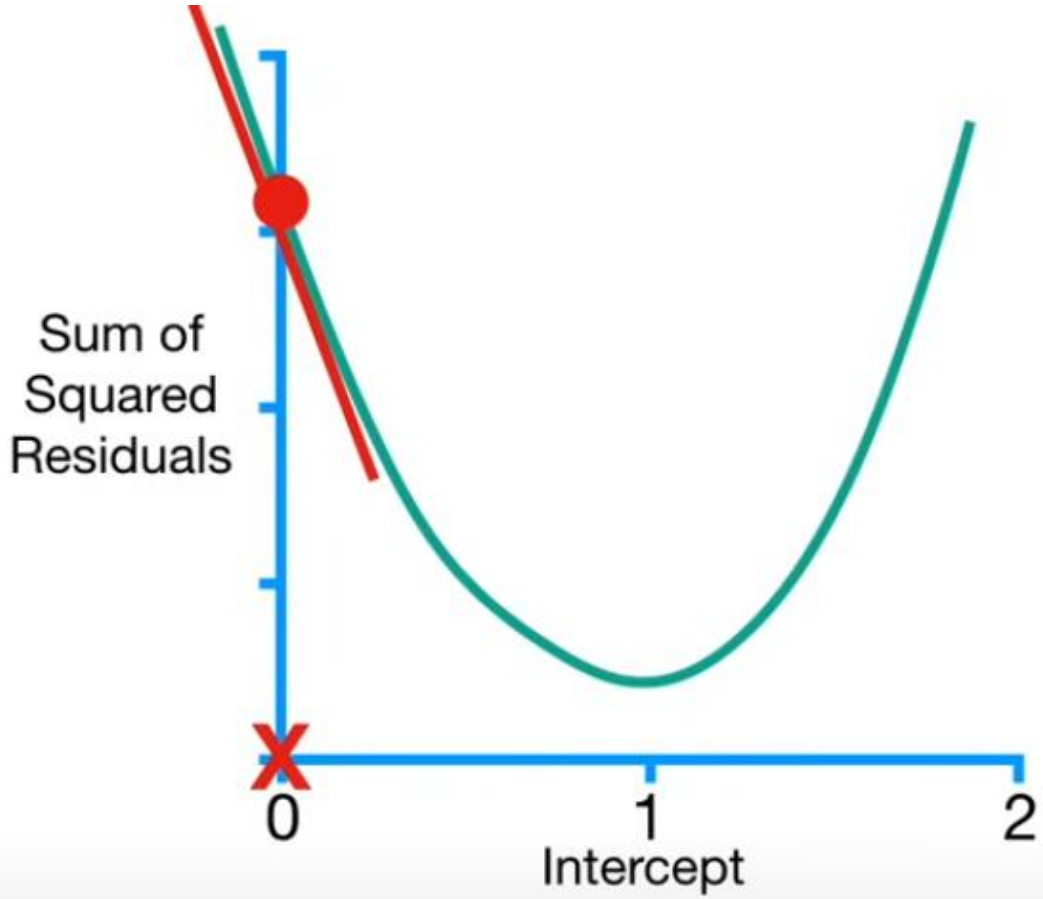
$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

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

$$= -5.7$$

Step Size = -5.7 × 0.1

…by a small number called
The Learning Rate.



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

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

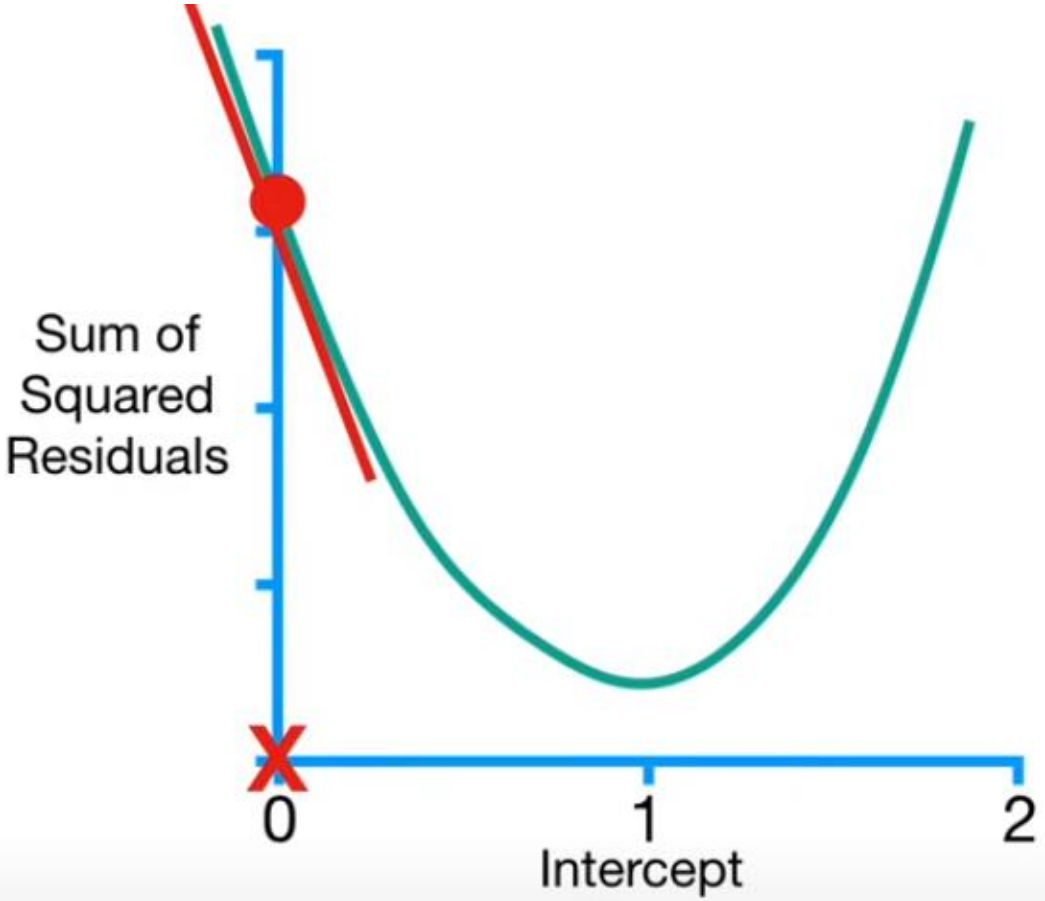
$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

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

$$= -5.7$$

Step Size = -5.7 × 0.1 = -0.57

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



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

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

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

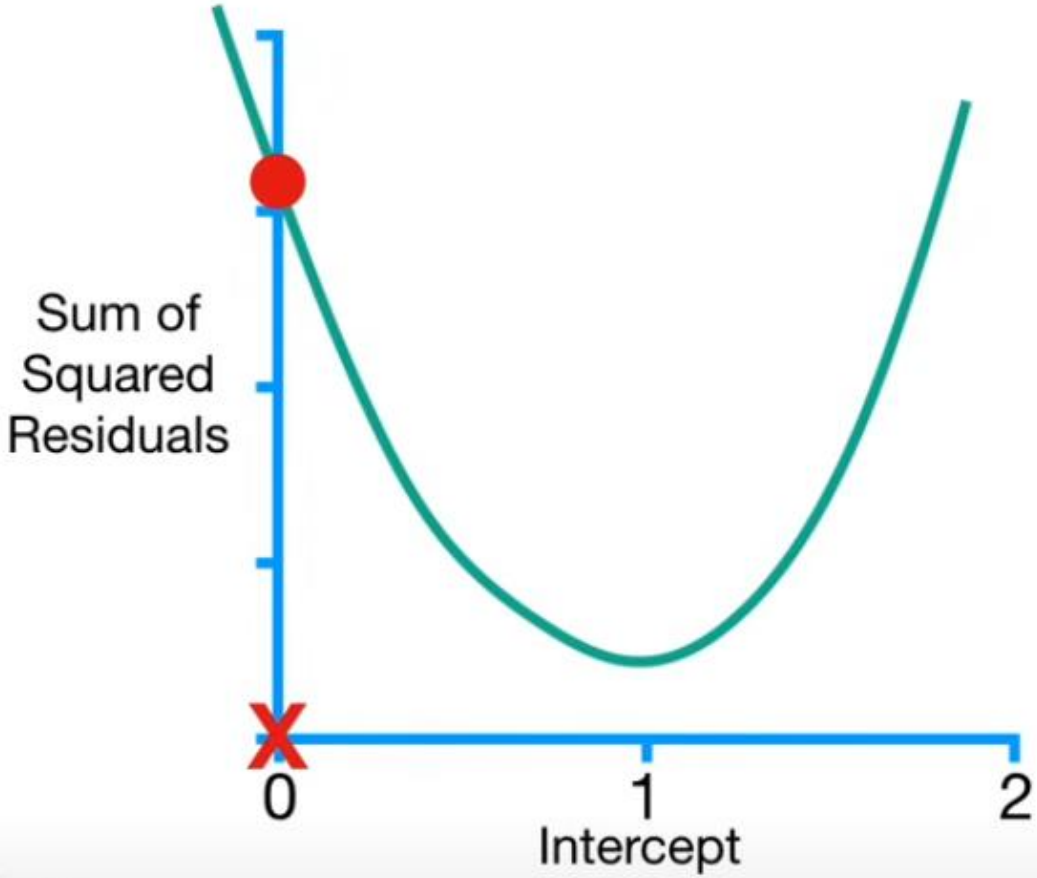
$$+ -2(\mathbf{3.2} - (0 + 0.64 \times \mathbf{2.9}))$$

$$= -5.7$$

Step Size = -5.7 × 0.1 = -0.57

New Intercept = ←

With the **Step Size**, we can calculate a **New Intercept**.



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

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

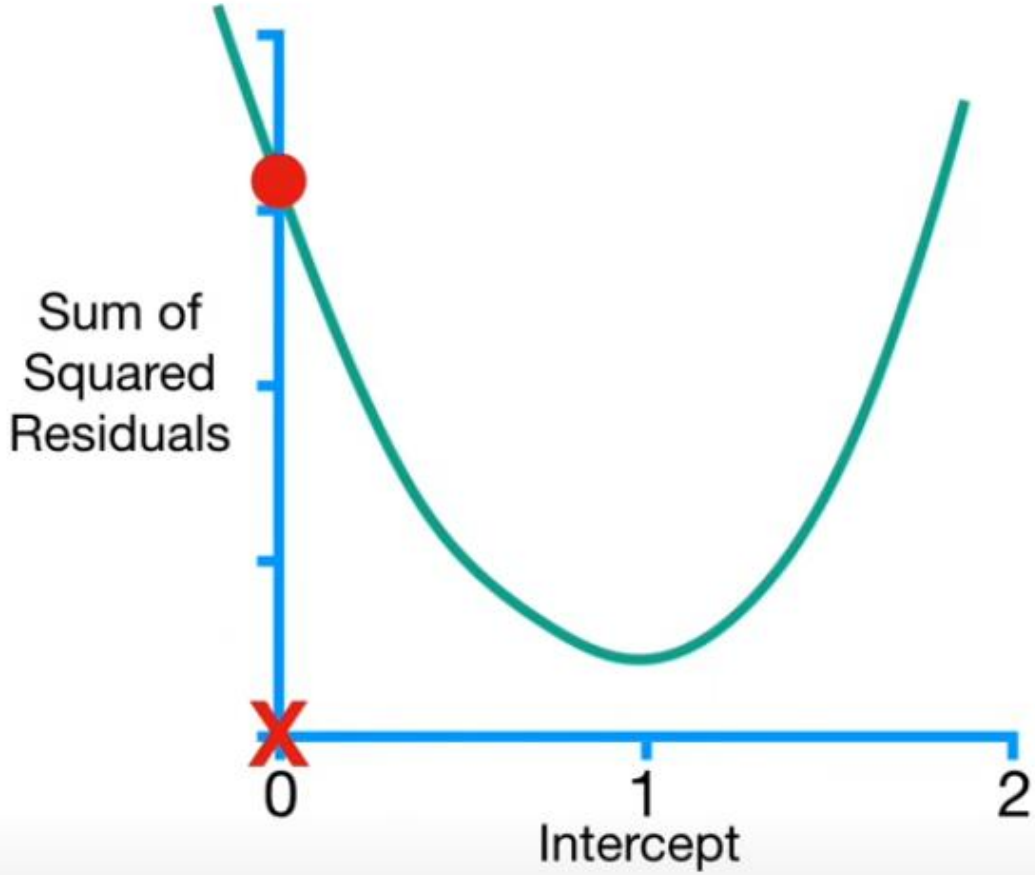
$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

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

$$= -5.7$$

Step Size = -5.7 x 0.1 = -0.57

New Intercept = 0 - **Step Size**



So we plug in the numbers…

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

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

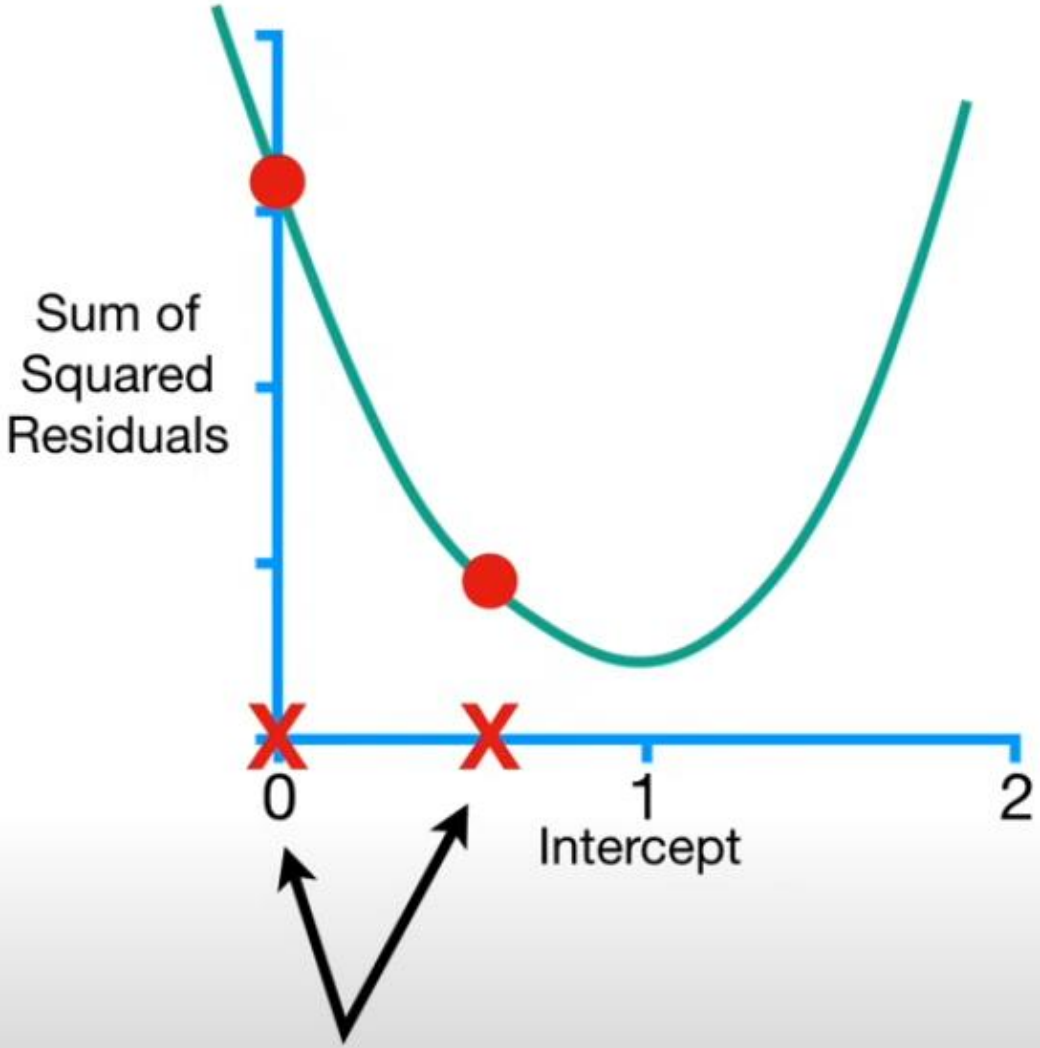
$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

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

$$= -5.7$$

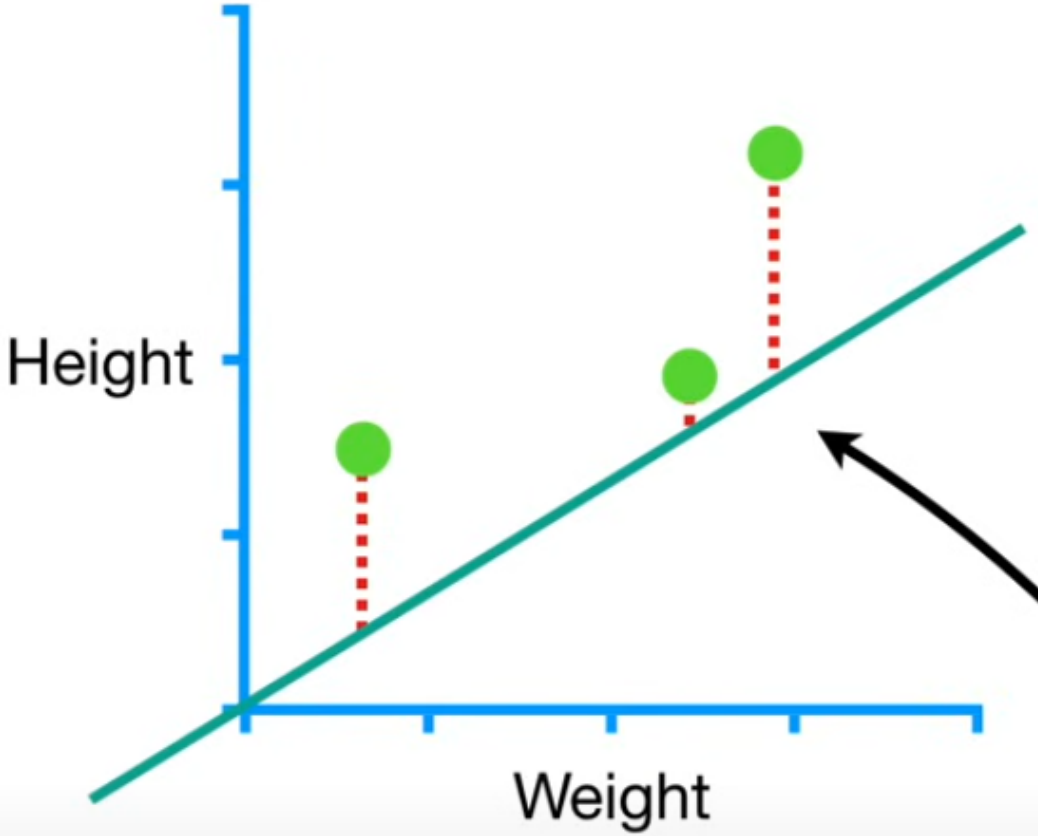
Step Size $= -5.7 \times 0.1 = -0.57$

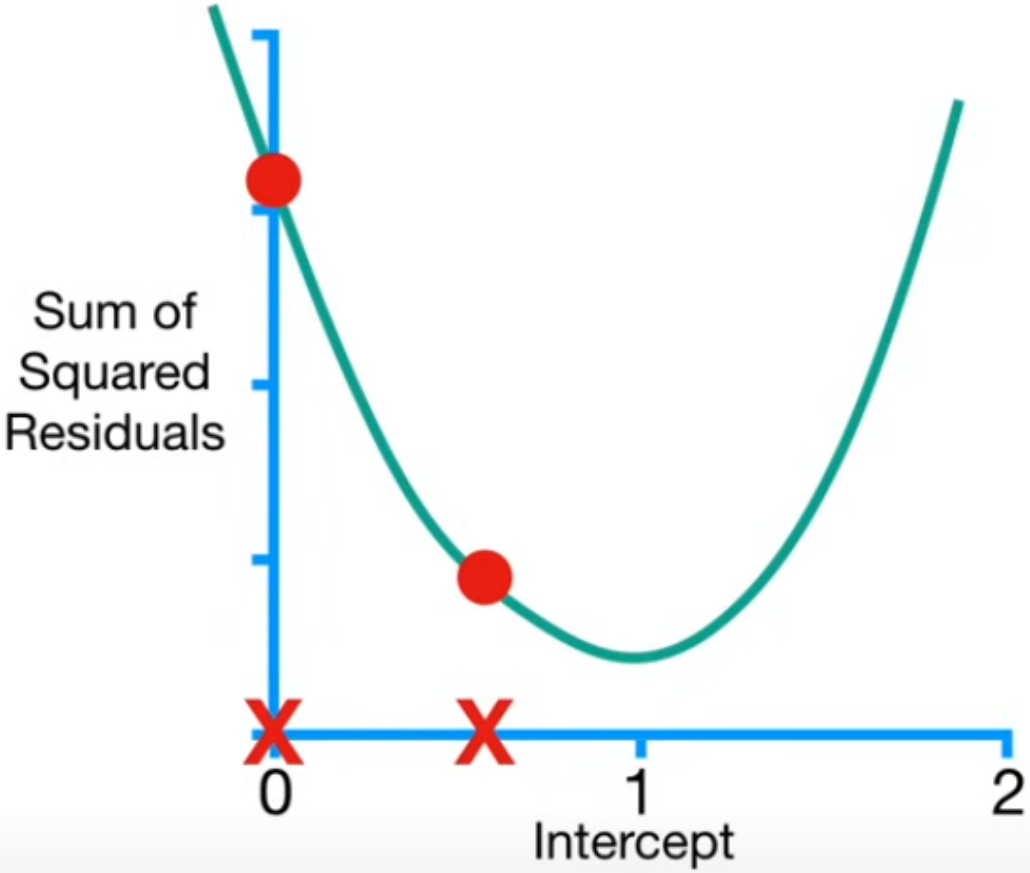
New Intercept $= 0 - (-0.57) = \mathbf{0.57}$



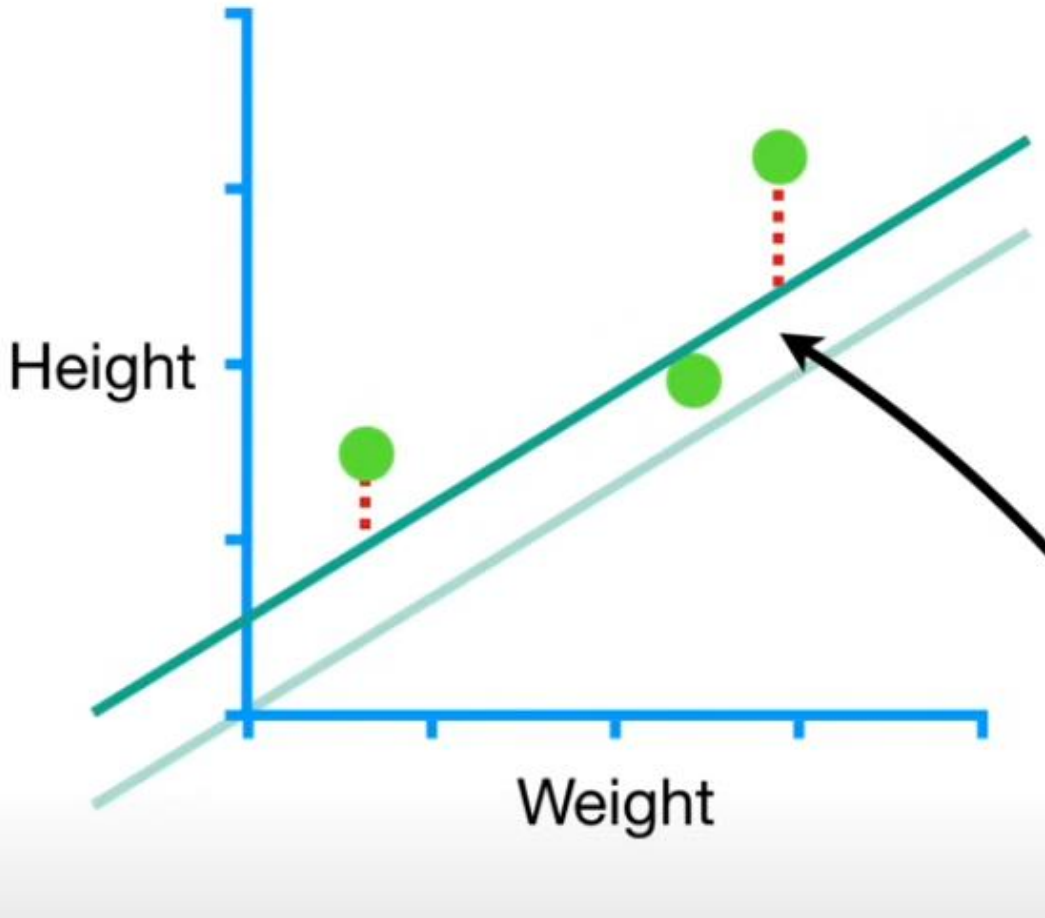
BAM!!!

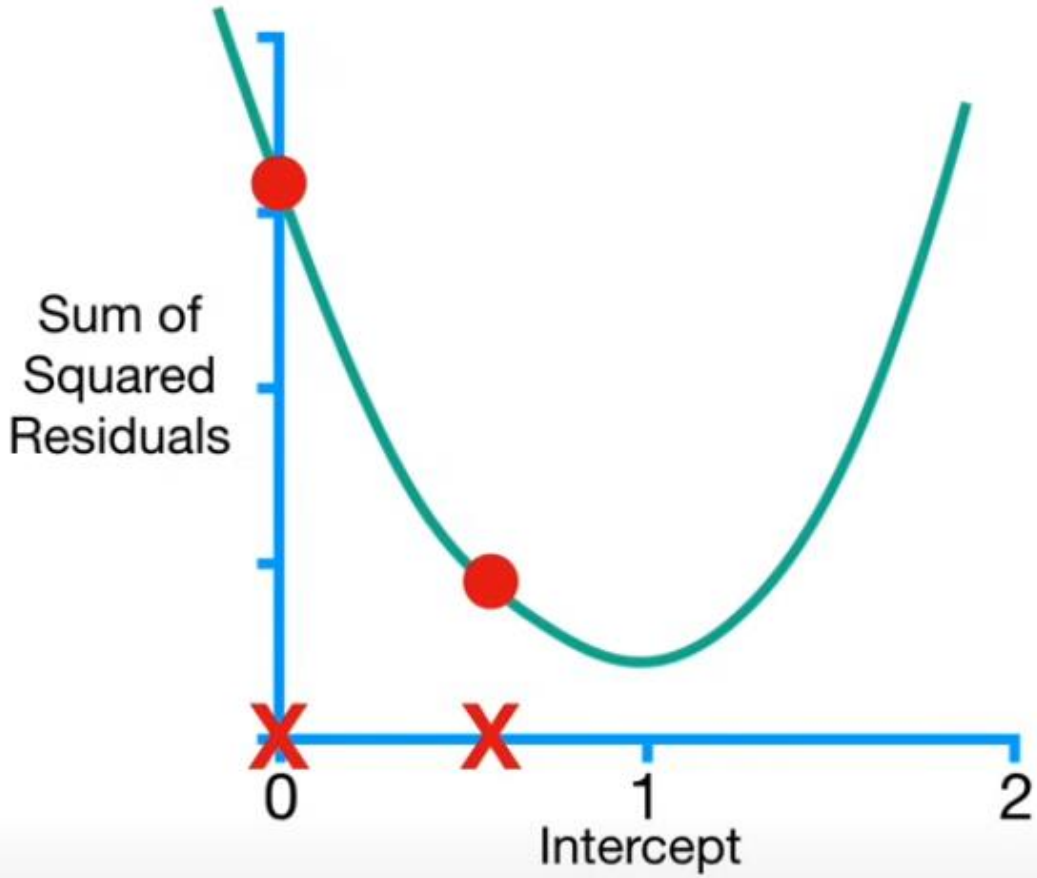
In one big step, we moved much closer to the optimal value for the **Intercept**.





Going back to the original data and the original line, with the **Intercept** = **0**…





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

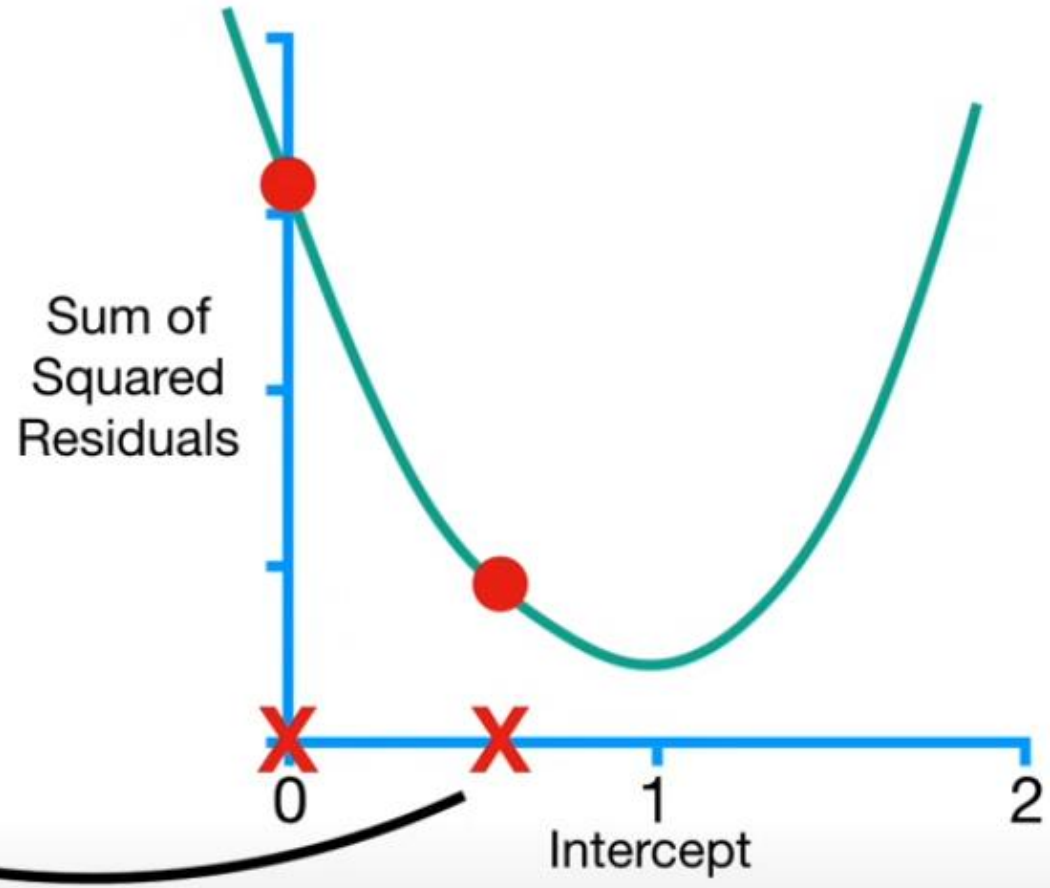
$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

-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 =

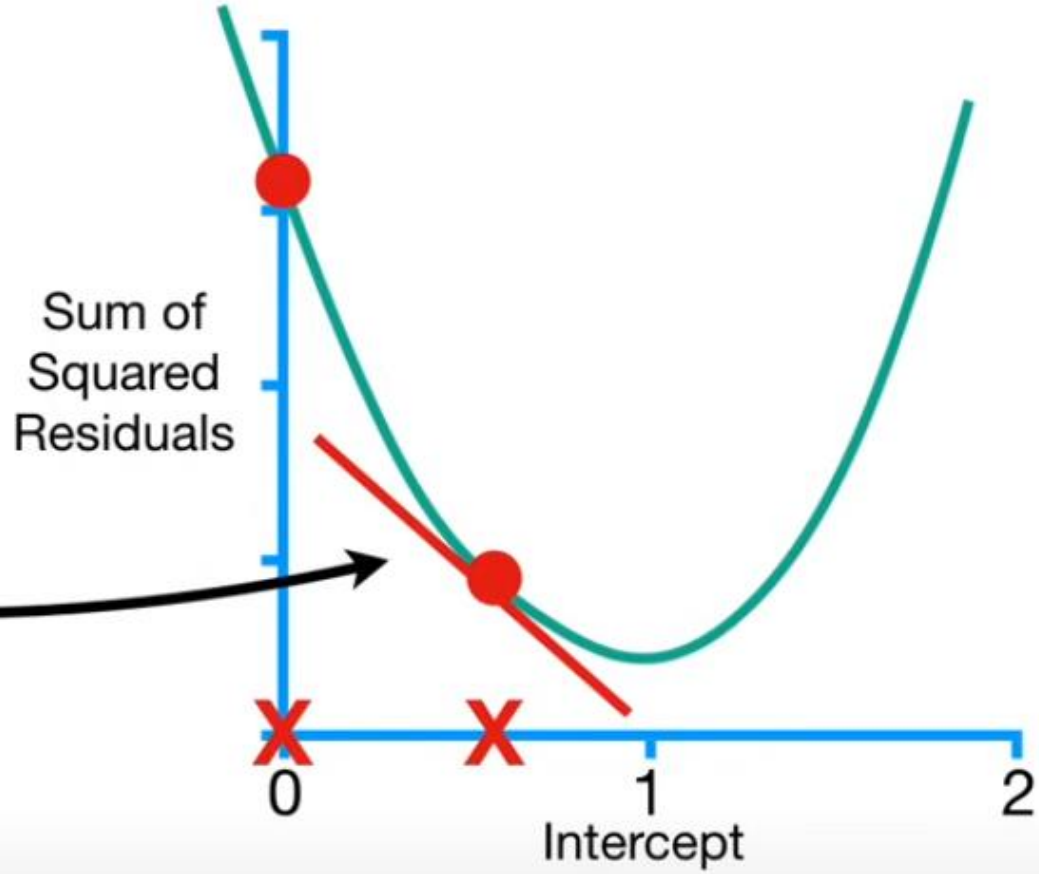
$$-2(\mathbf{1.4} - (0.57 + 0.64 \times \mathbf{0.5})$$

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

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

$$= \boxed{-2.3}$$

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



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

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

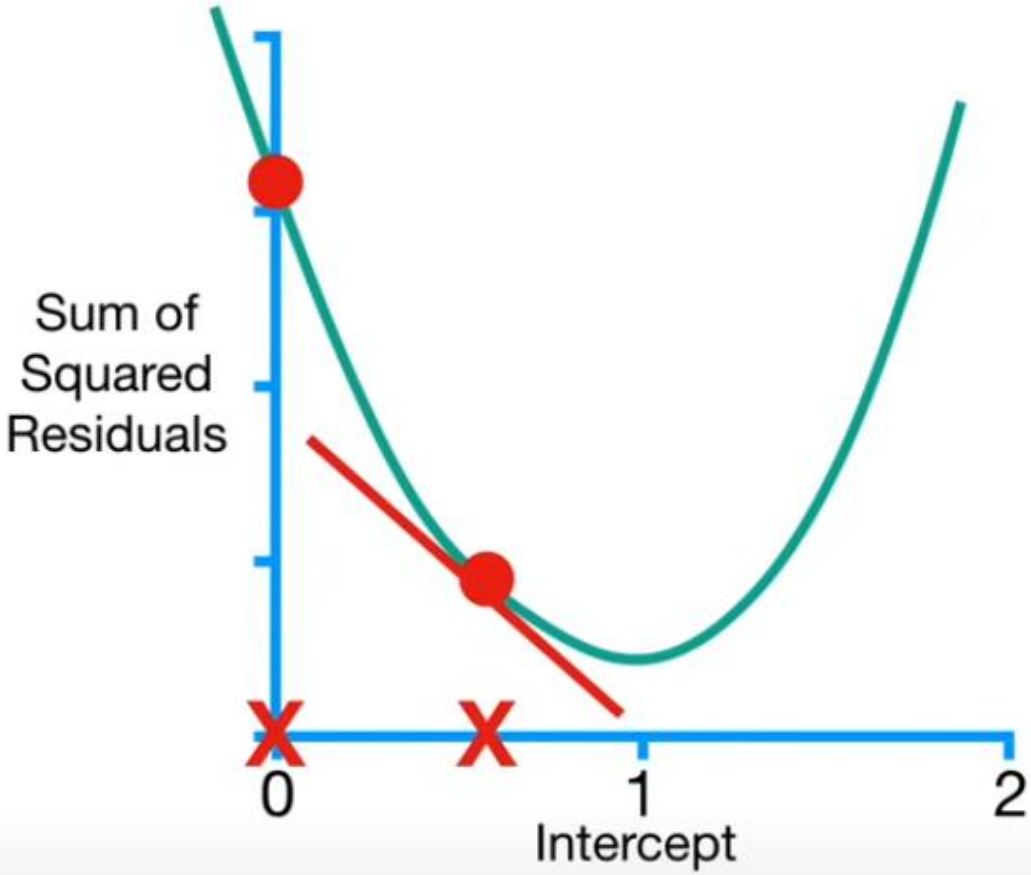
$$+ -2(\mathbf{1.9} - (0.57 + 0.64 \times \mathbf{2.3}))$$

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

$$= -2.3$$

Step Size = -2.3 × **Learning Rate**

…by plugging in **-2.3** for the **Slope**…



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

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

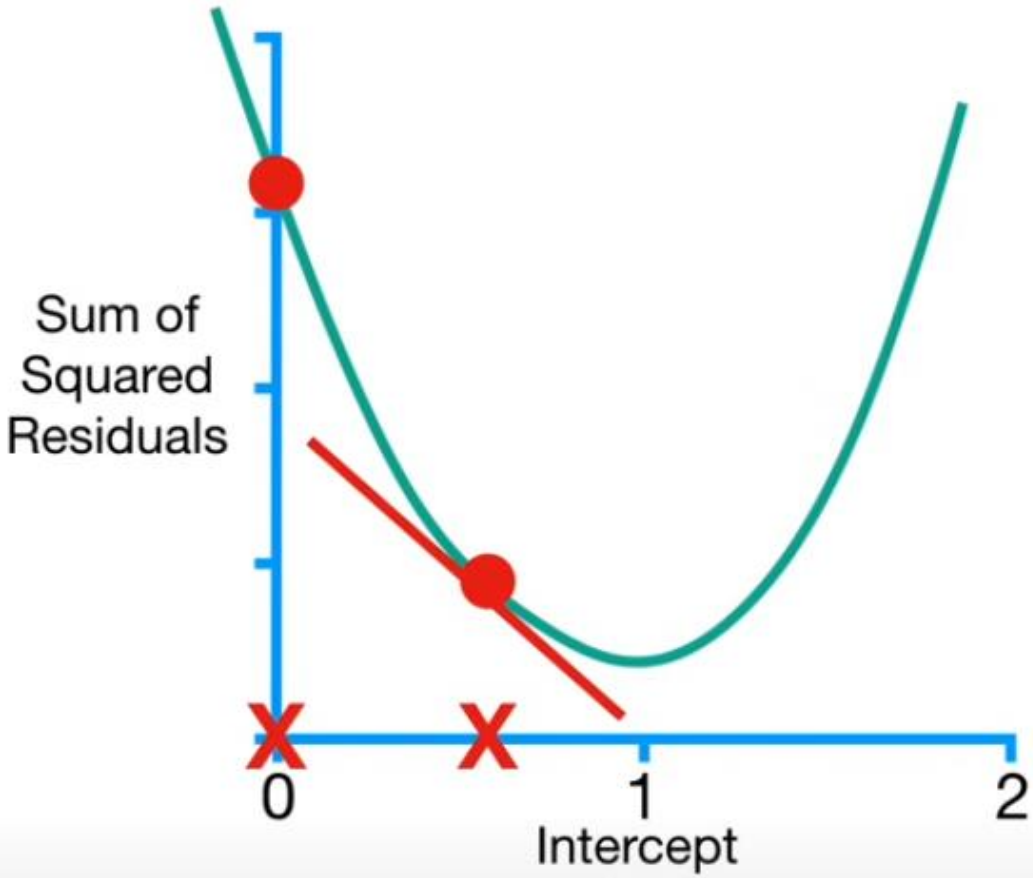
$$+ -2(\mathbf{1.9} - (0.57 + 0.64 \times \mathbf{2.3}))$$

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

$$= -2.3$$

Step Size = -2.3 × 0.1 = **-0.23**

Ultimately, the **Step Size** is **-0.23**…



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

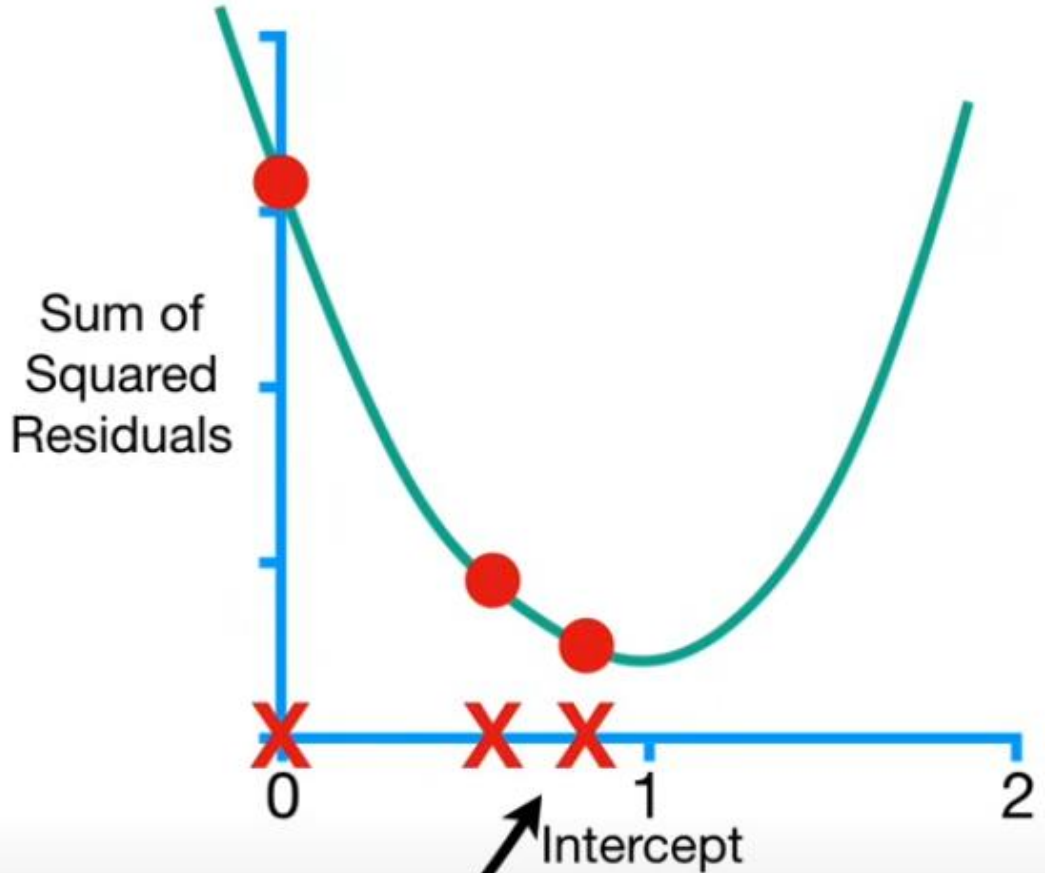
$$-2(\mathbf{1.4} - (0.57 + 0.64 \times \mathbf{0.5})$$

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

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

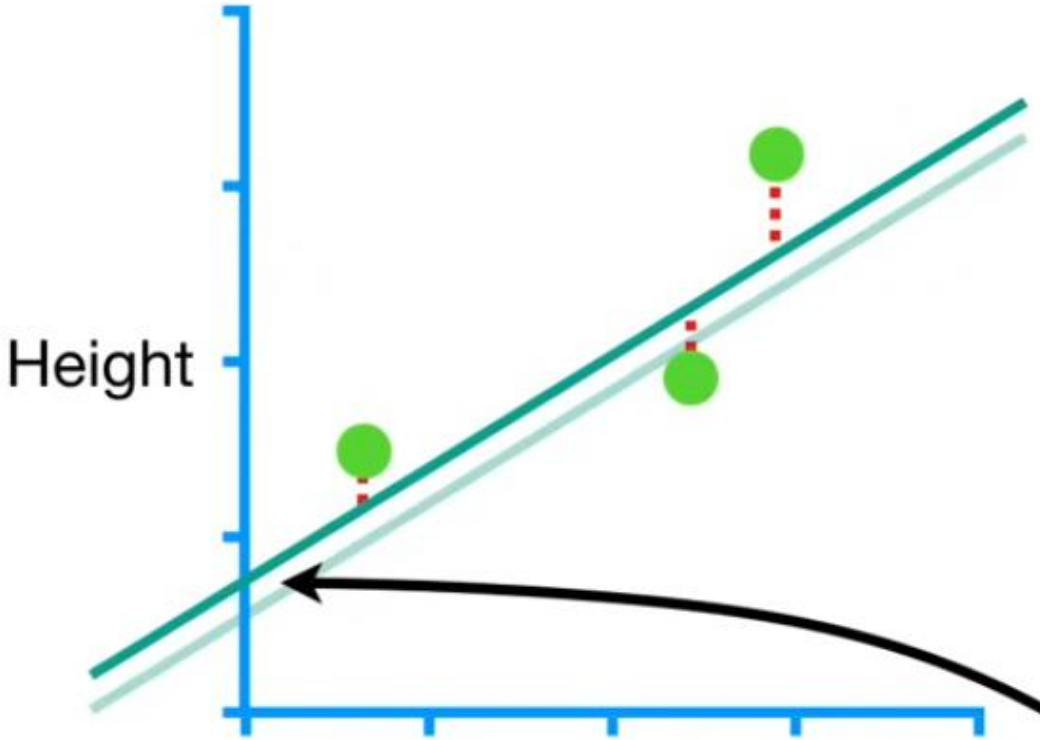
$$= -2.3$$

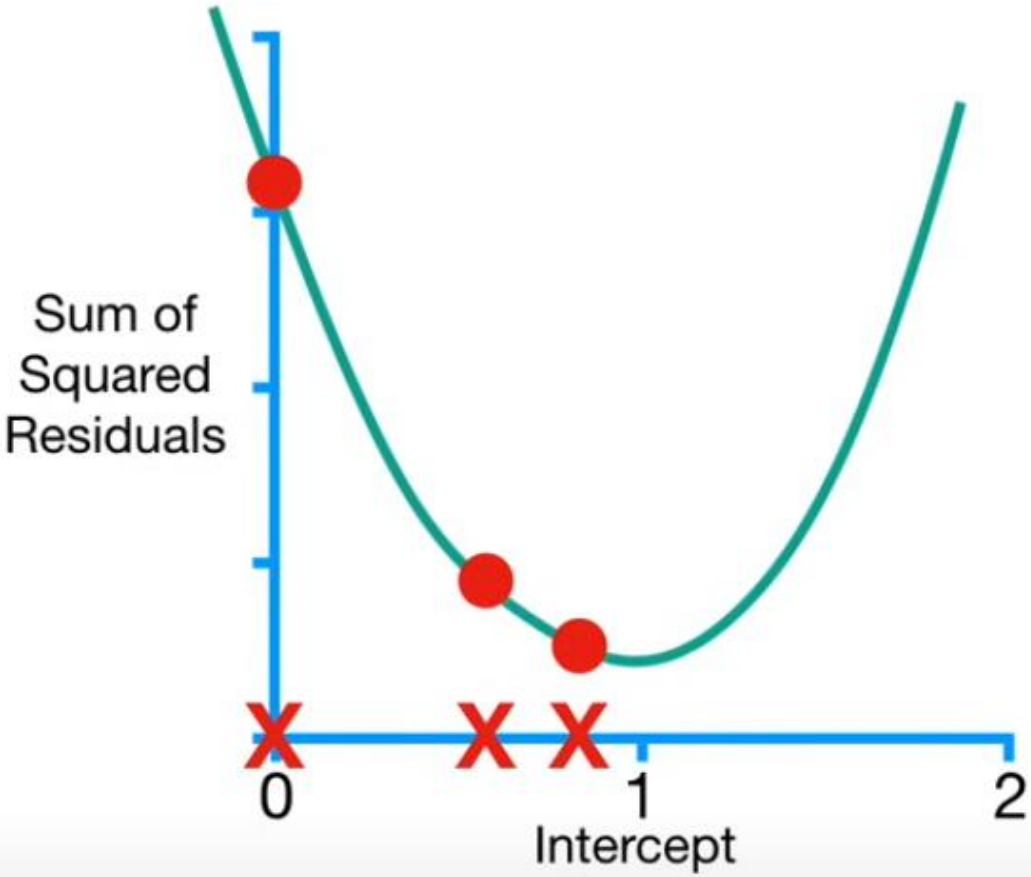
Step Size $= -2.3 \times 0.1 = \mathbf{-0.23}$



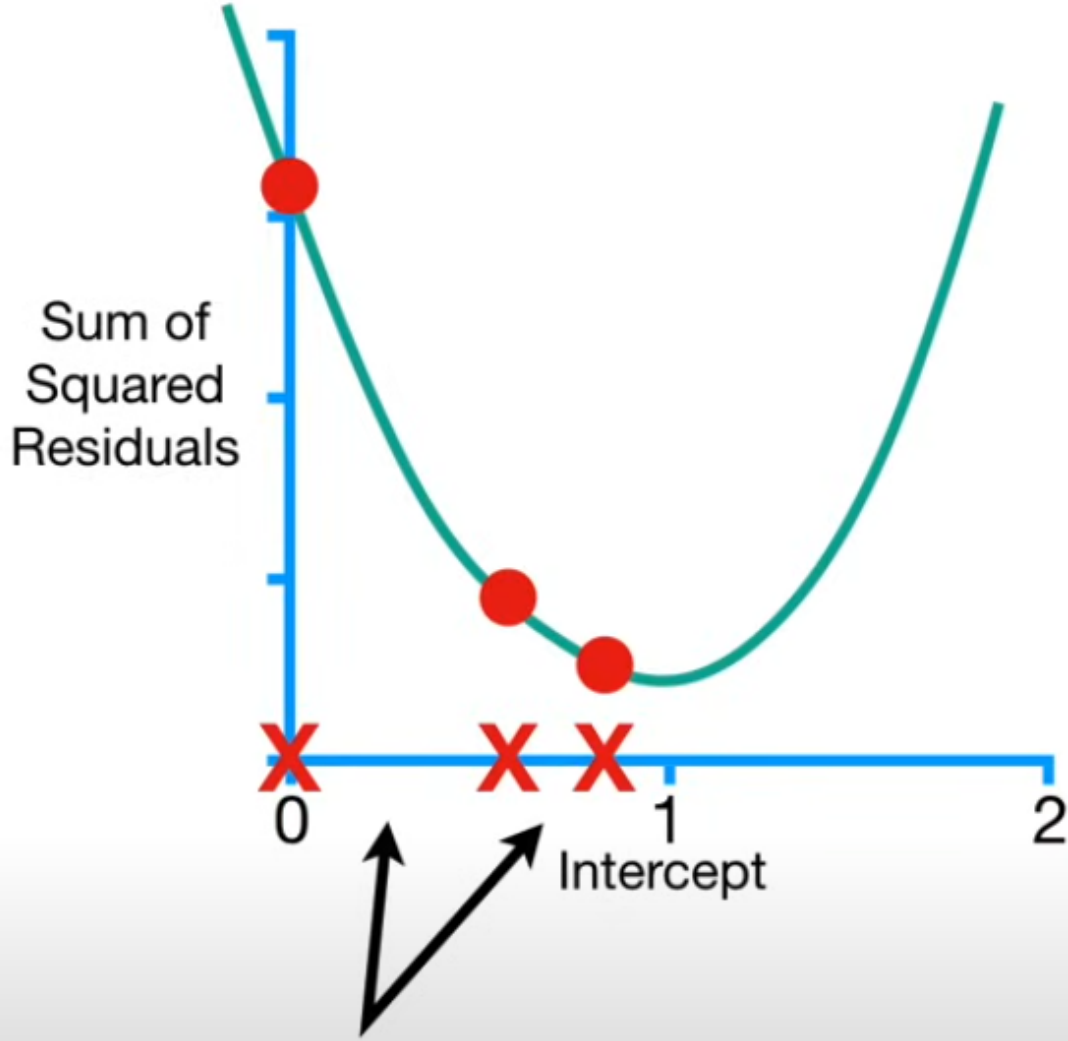
New Intercept $= 0.57 - (-0.23) = \mathbf{0.8}$

…and the **New Intercept** = **0.8**





…to when the
Intercept = 0.8



Notice that the first step was relatively large compared to the second step.

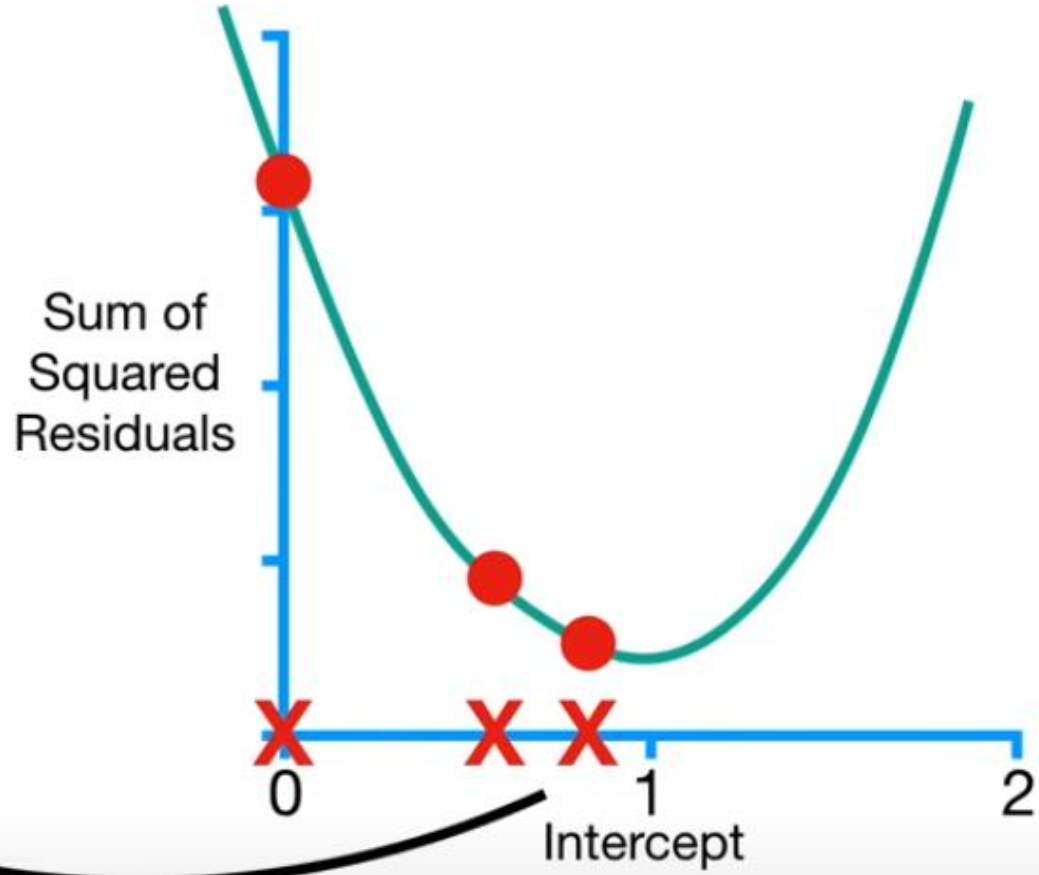
$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

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

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

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

Now let's calculate the derivative at the **New Intercept** (**0.8**)…



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

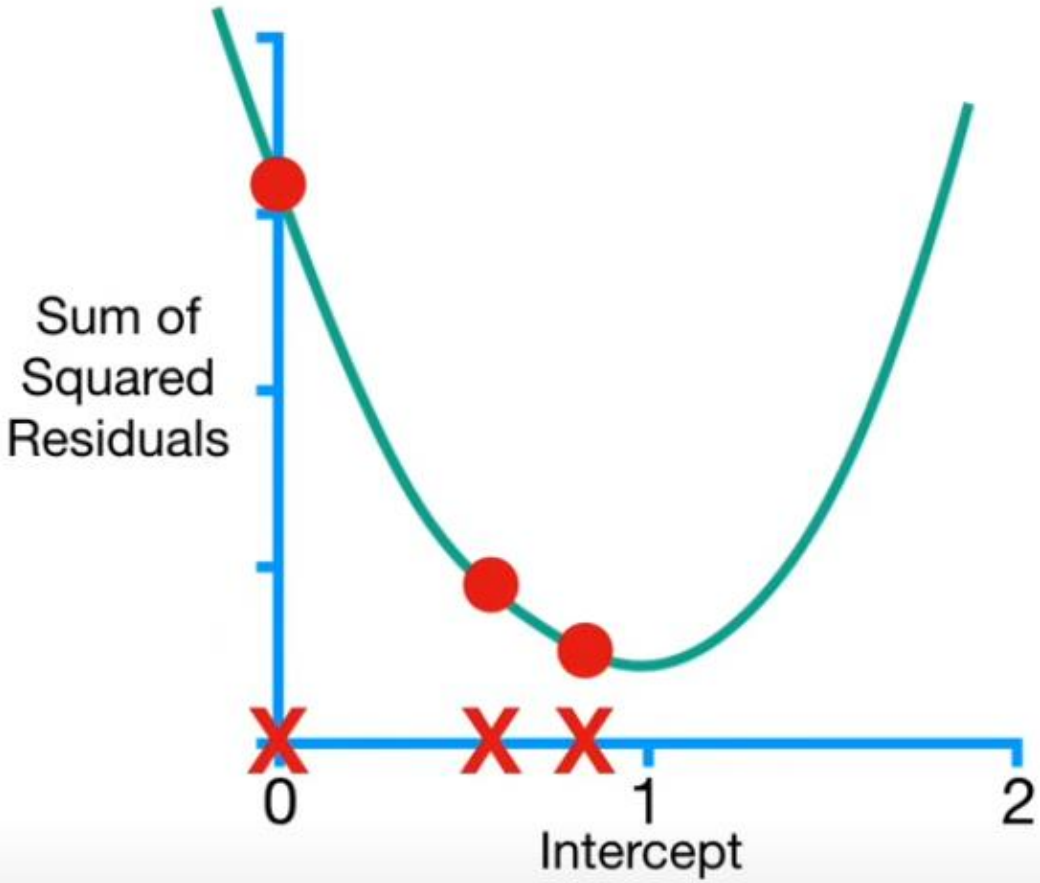
$$-2(\mathbf{1.4} - (0.8 + 0.64 \times \mathbf{0.5})$$

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

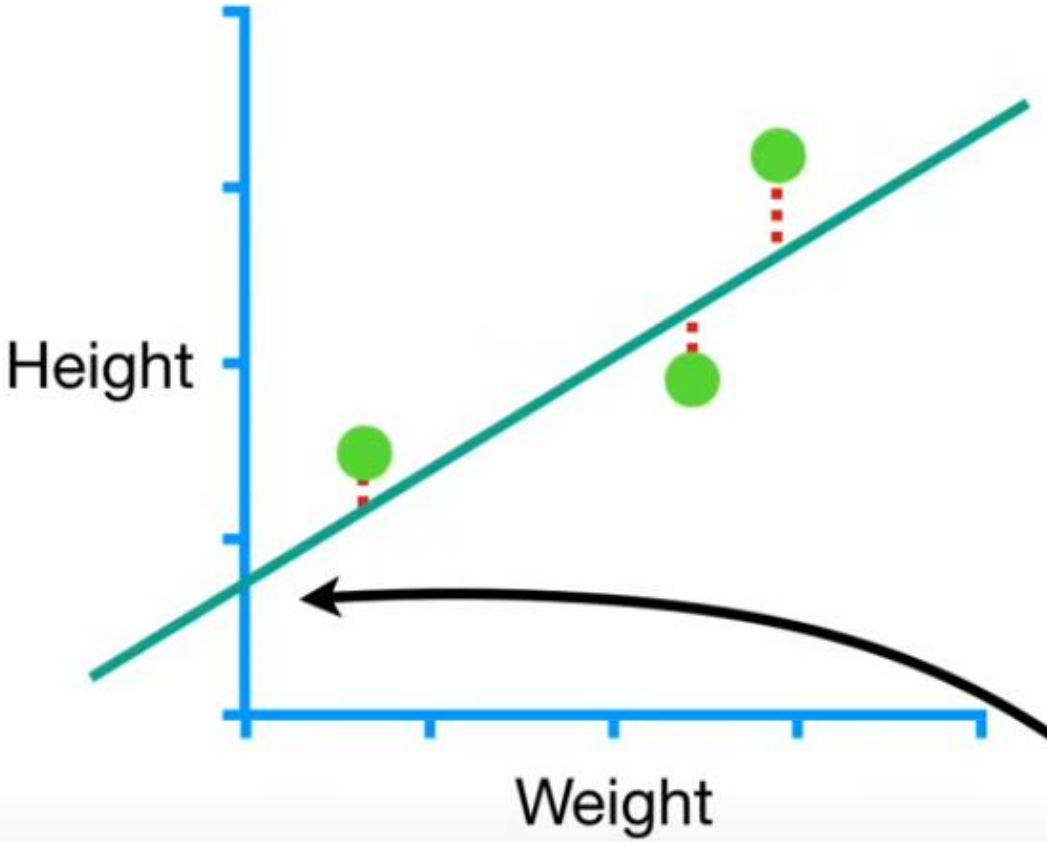
$$+ -2(\mathbf{3.2} - (0.8 + 0.64 \times \mathbf{2.9}))$$

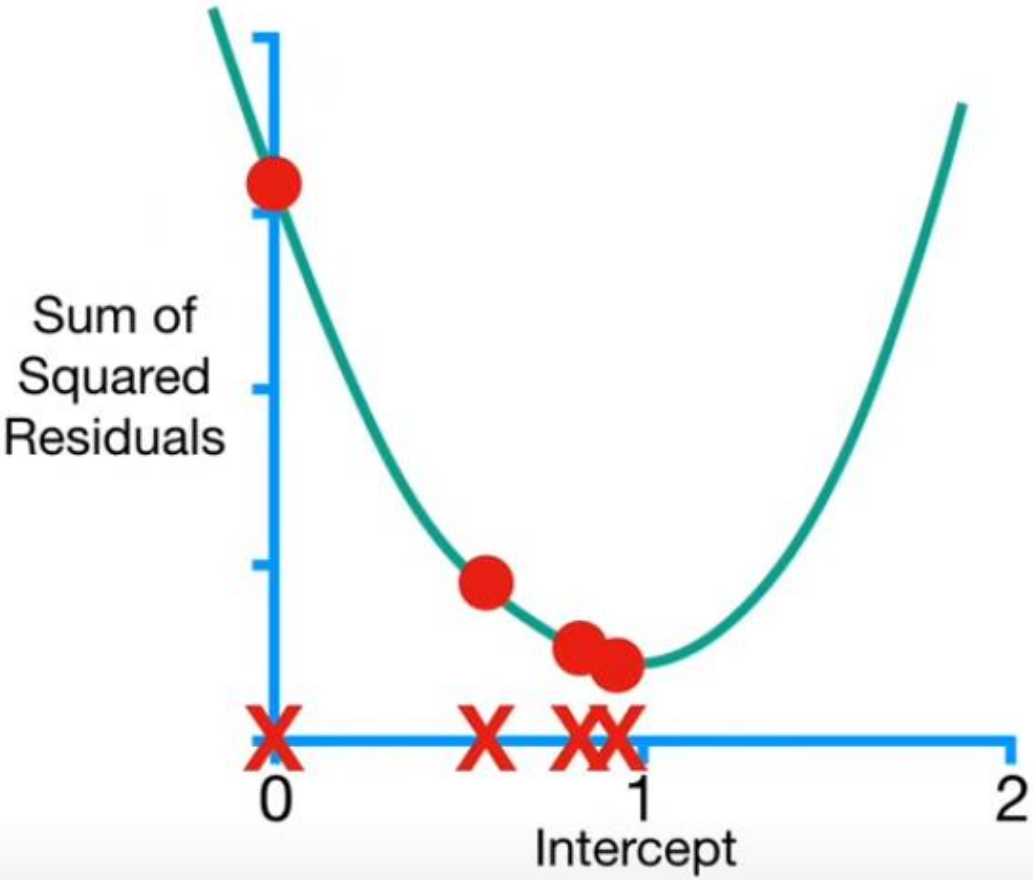
$$= \mathbf{-0.9}$$

Step Size = -0.9 × 0.1 = **-0.09**

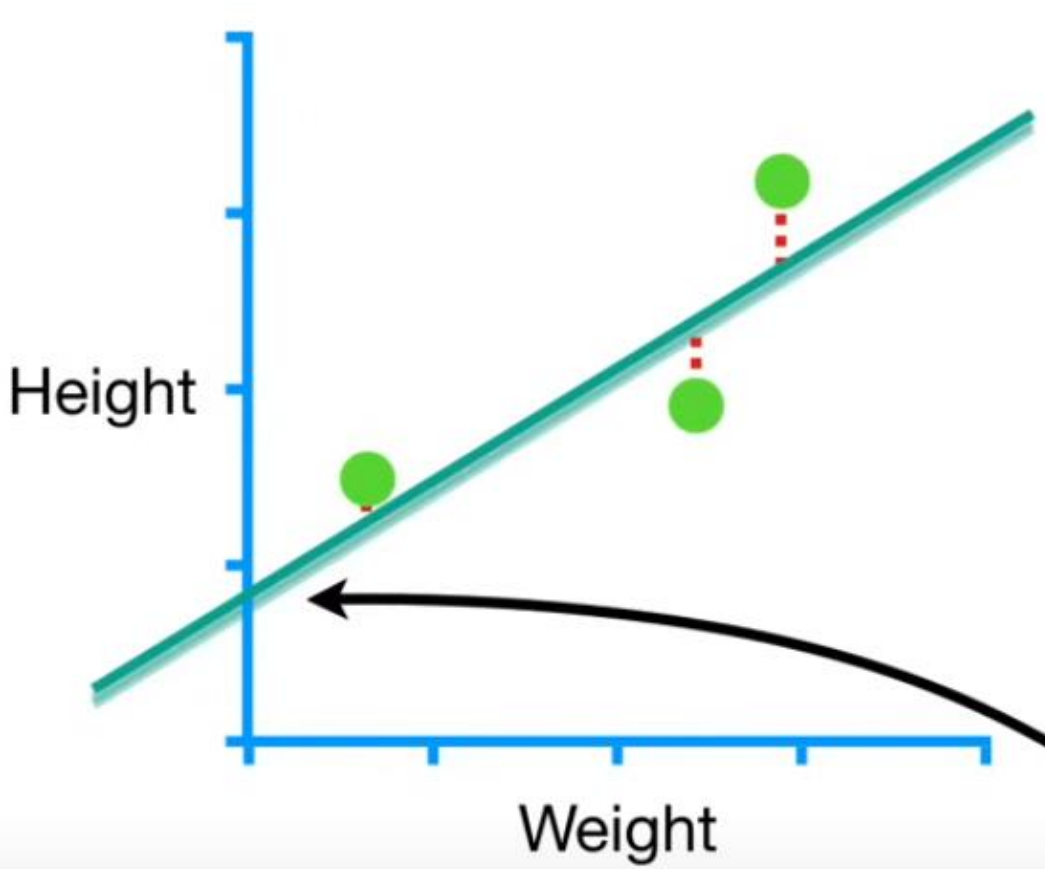


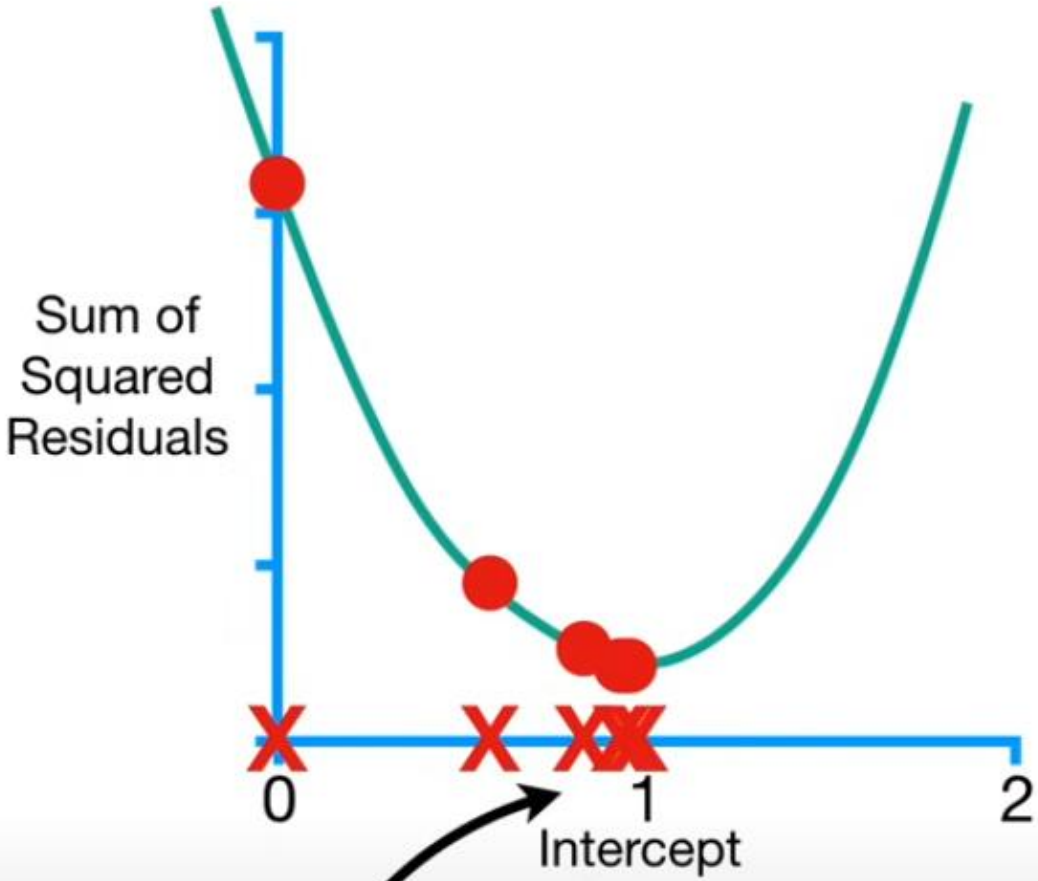
The **Step Size** = **-0.09**…



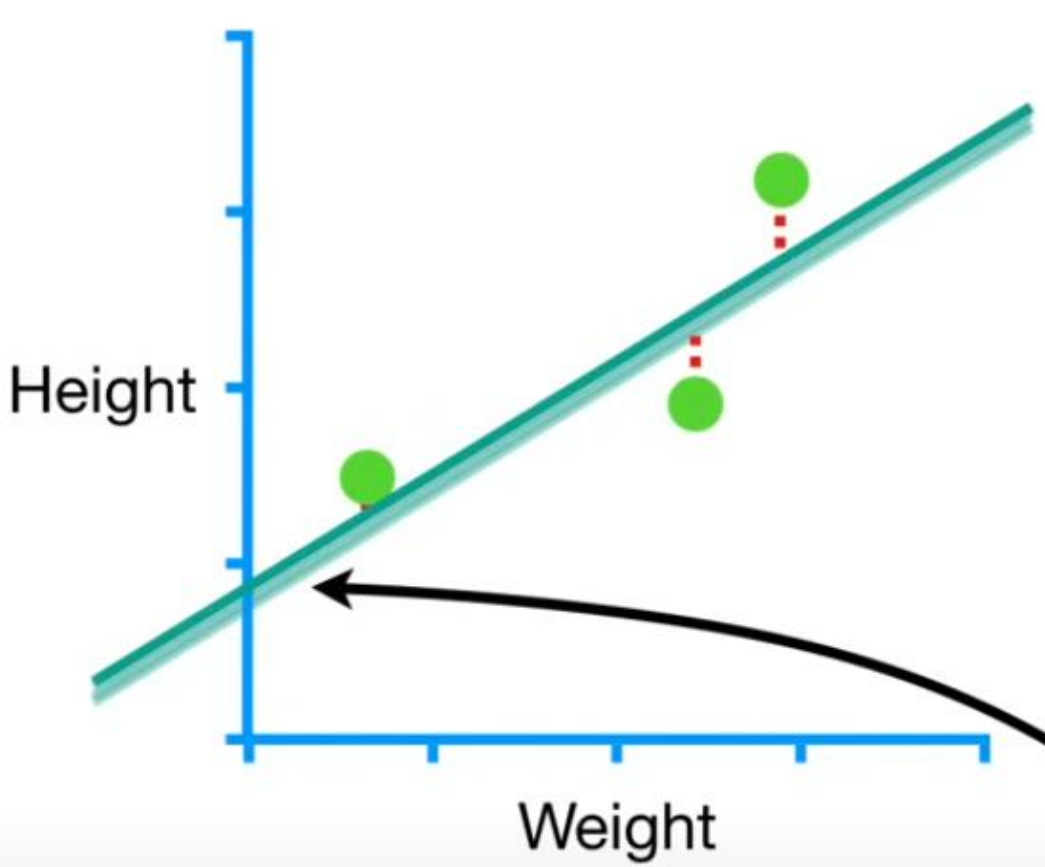


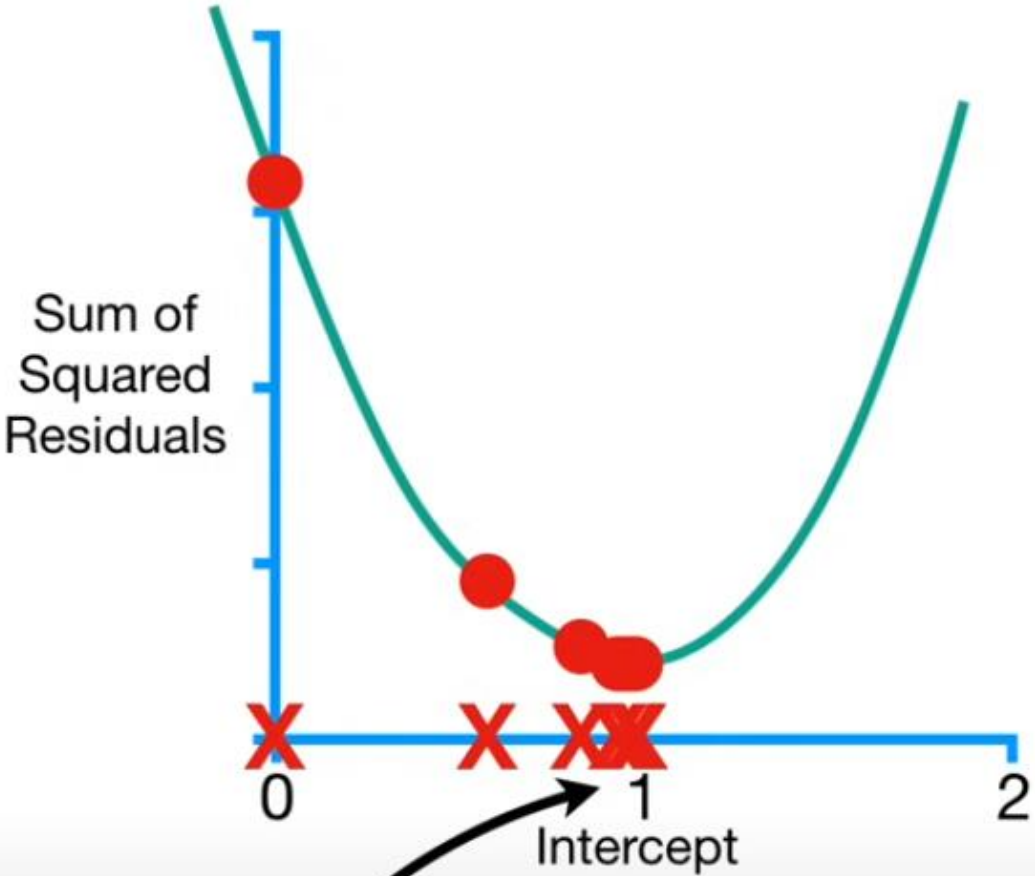
Now we increase the
Intercept from **0.8**…



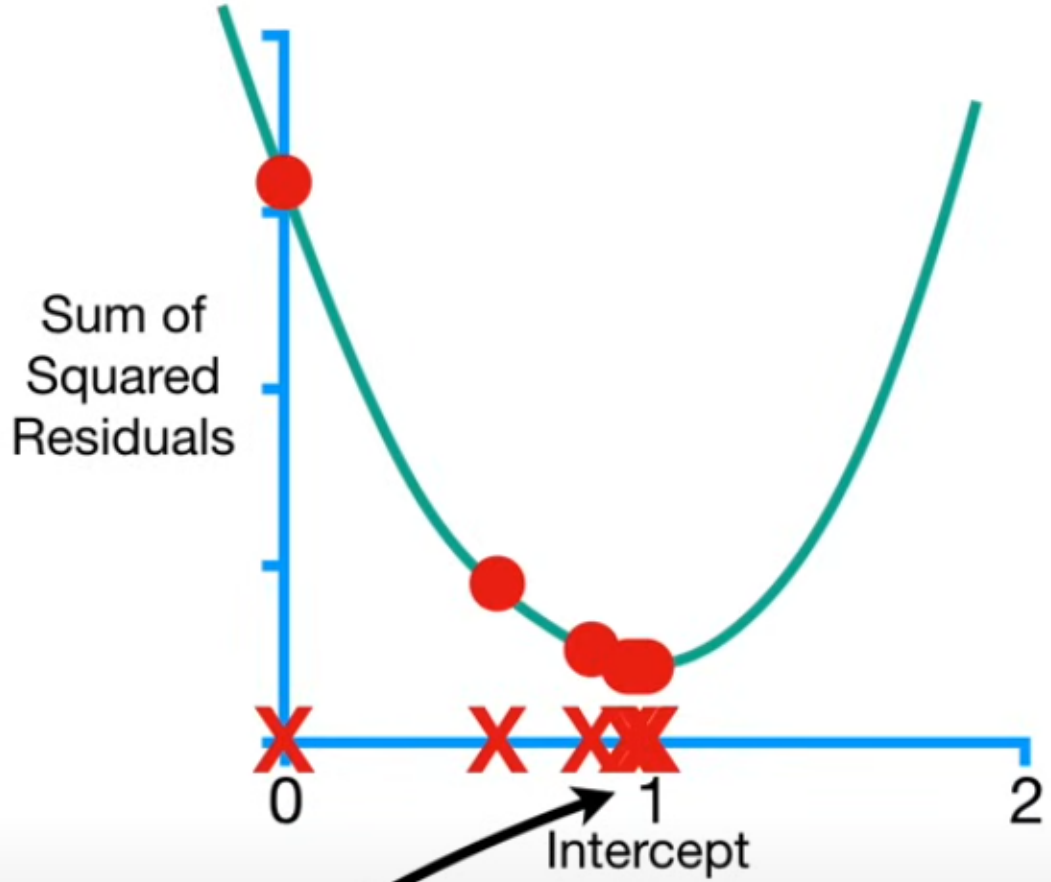


Then we take another step and
the **New Intercept** = **0.92**…



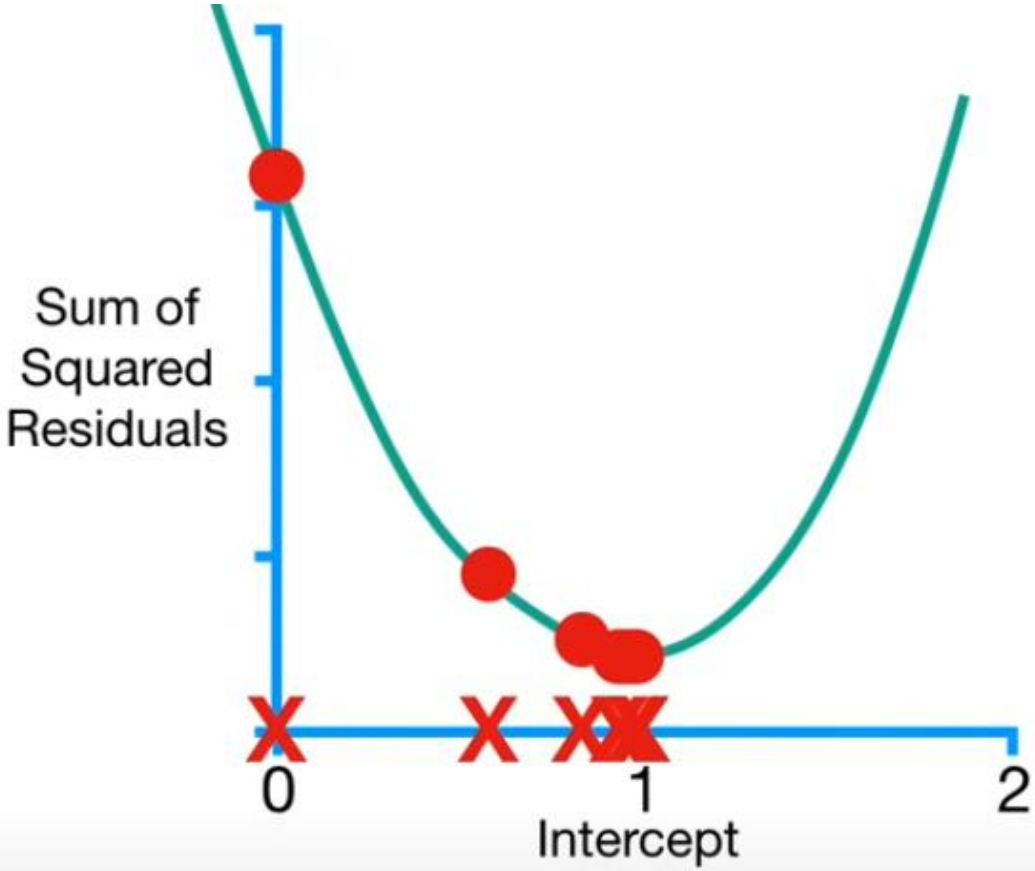


…and then we take another step and the **New Intercept = 0.94**…



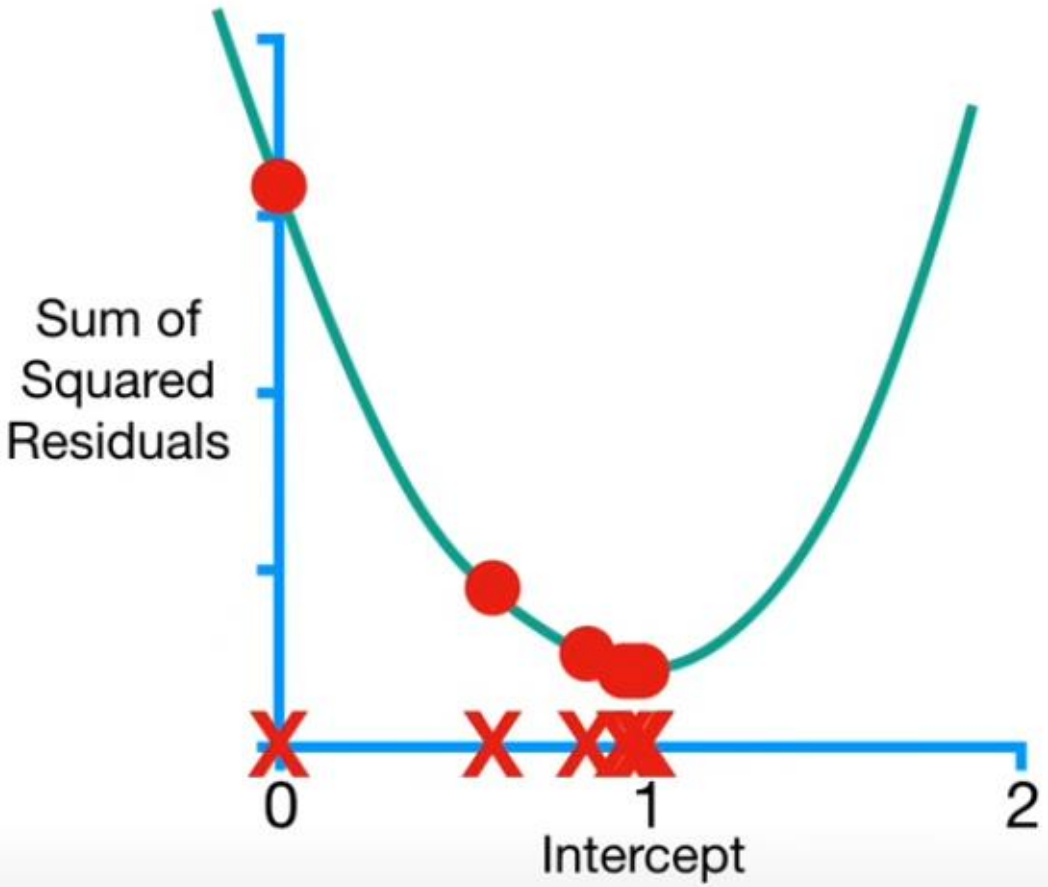
Notice how each step gets smaller and smaller the closer we get to the bottom of the curve.

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



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

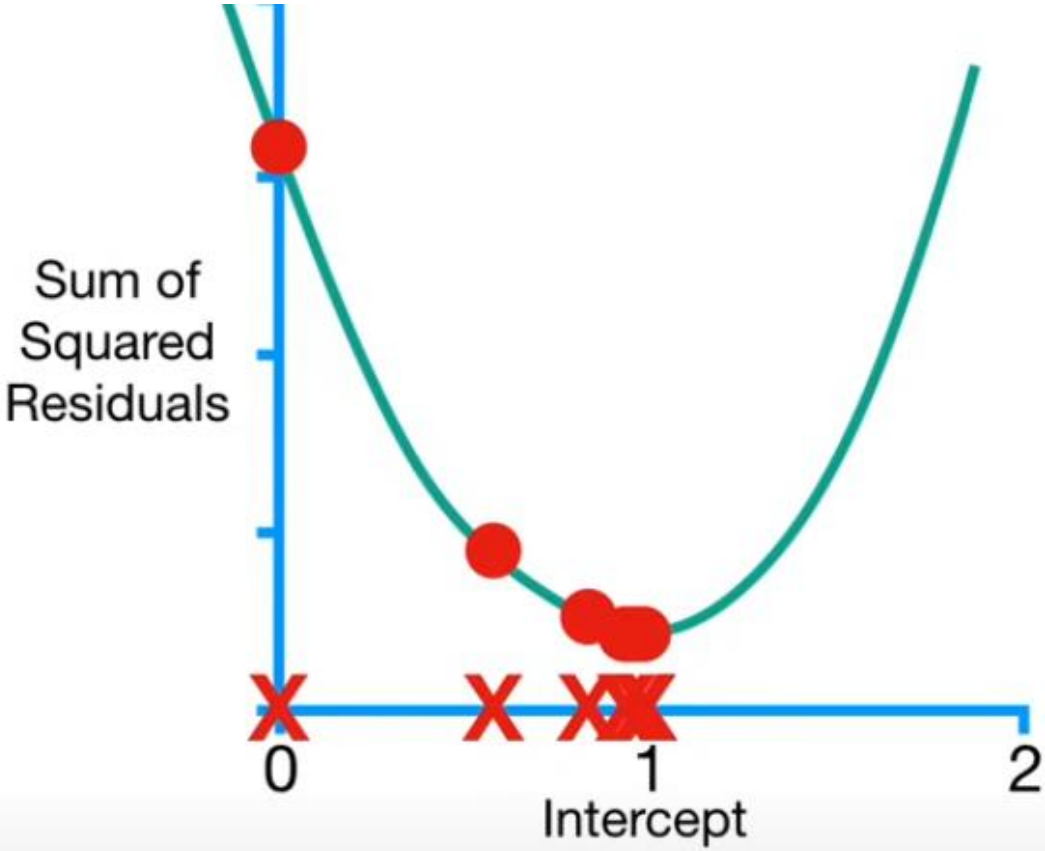
NOTE: The **Least Squares** estimate for the intercept is also **0.95**.



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

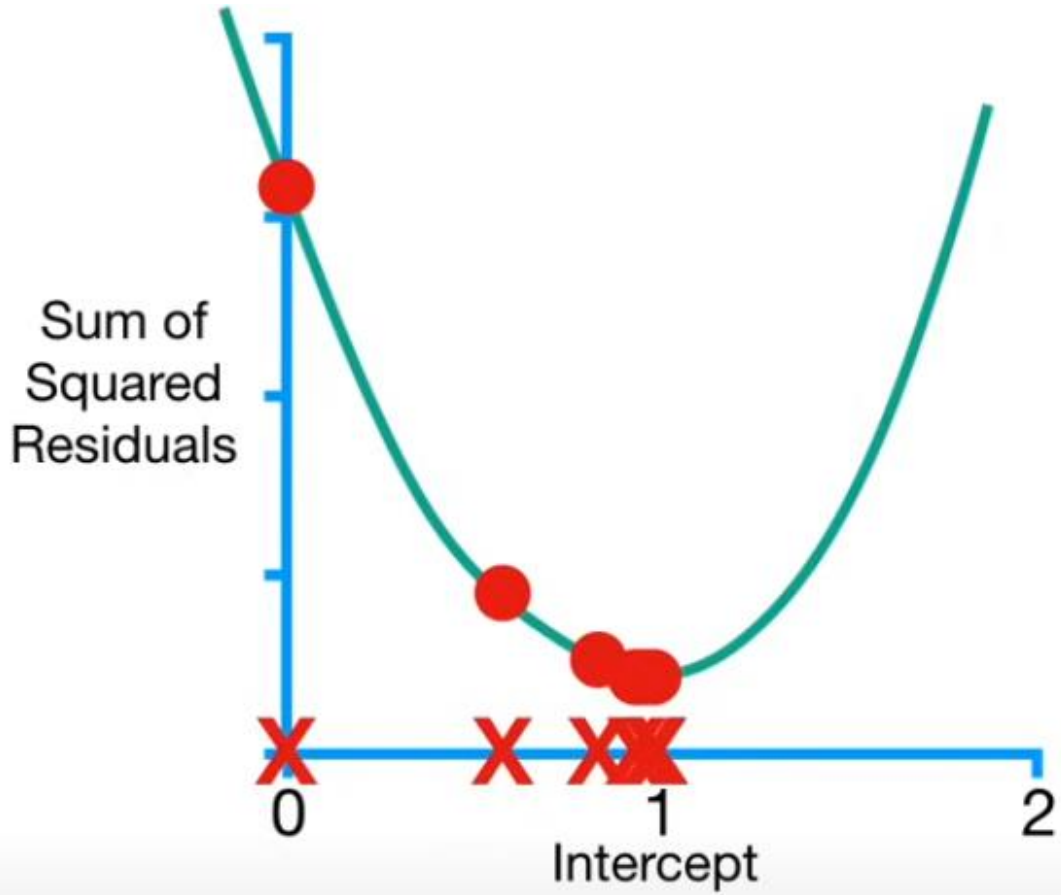
NOTE: The **Least Squares** estimate for the intercept is also **0.95**.

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?



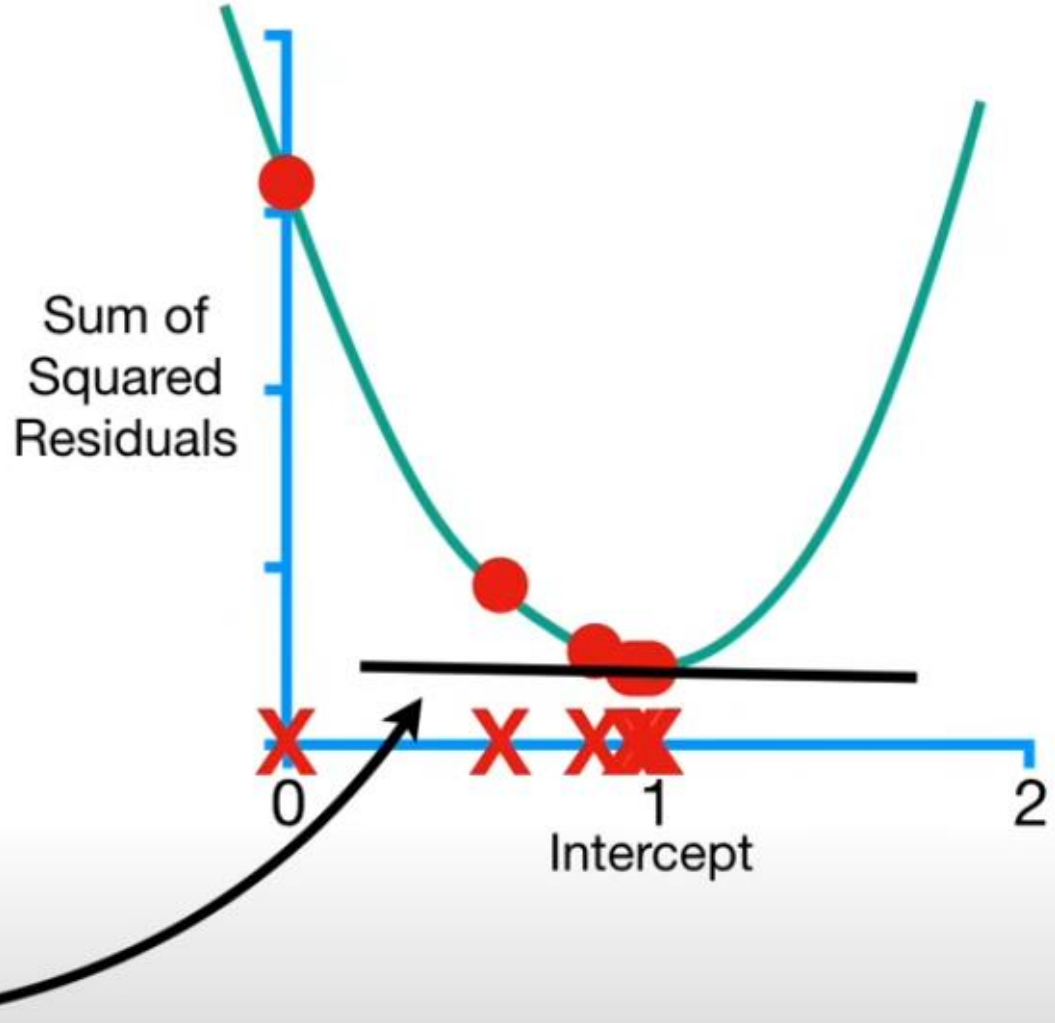
Gradient Descent stops when the **Step Size** is **Very Close To 0**.

Step Size = **Slope** x **Learning Rate**



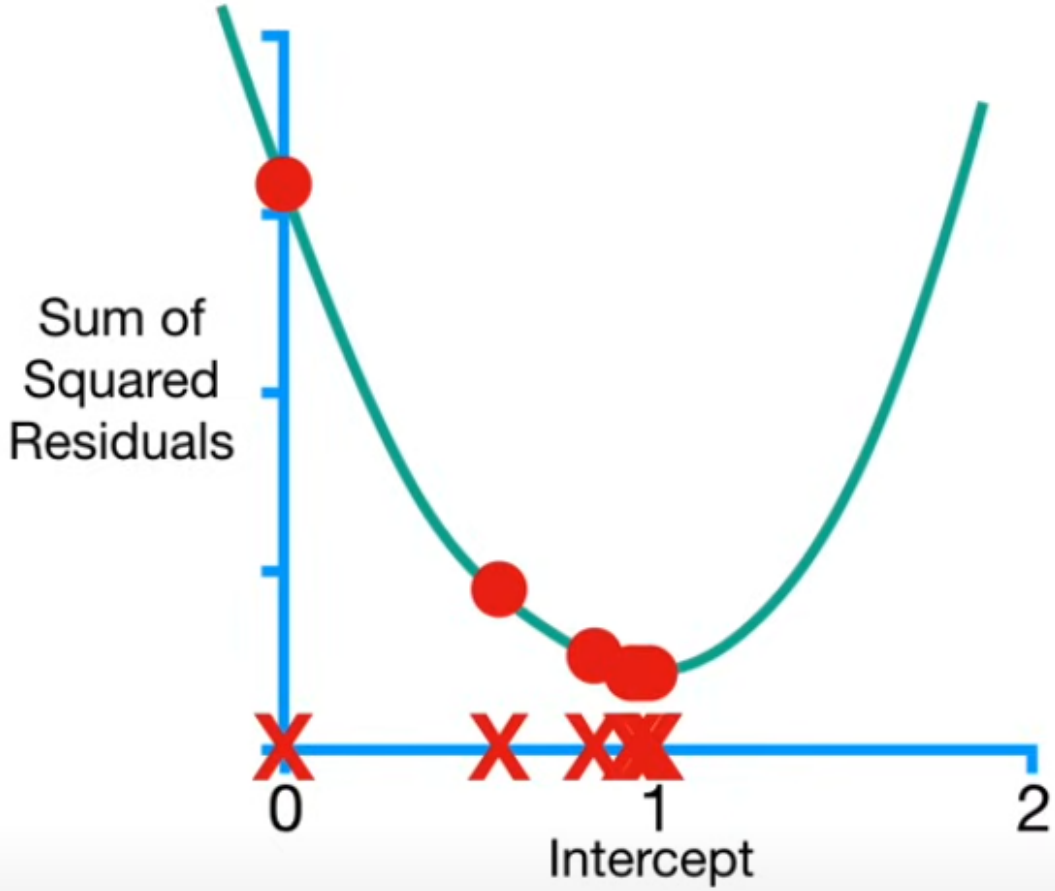
The **Step Size** will be **Very Close to 0** when the **Slope** is very close to **0**.

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



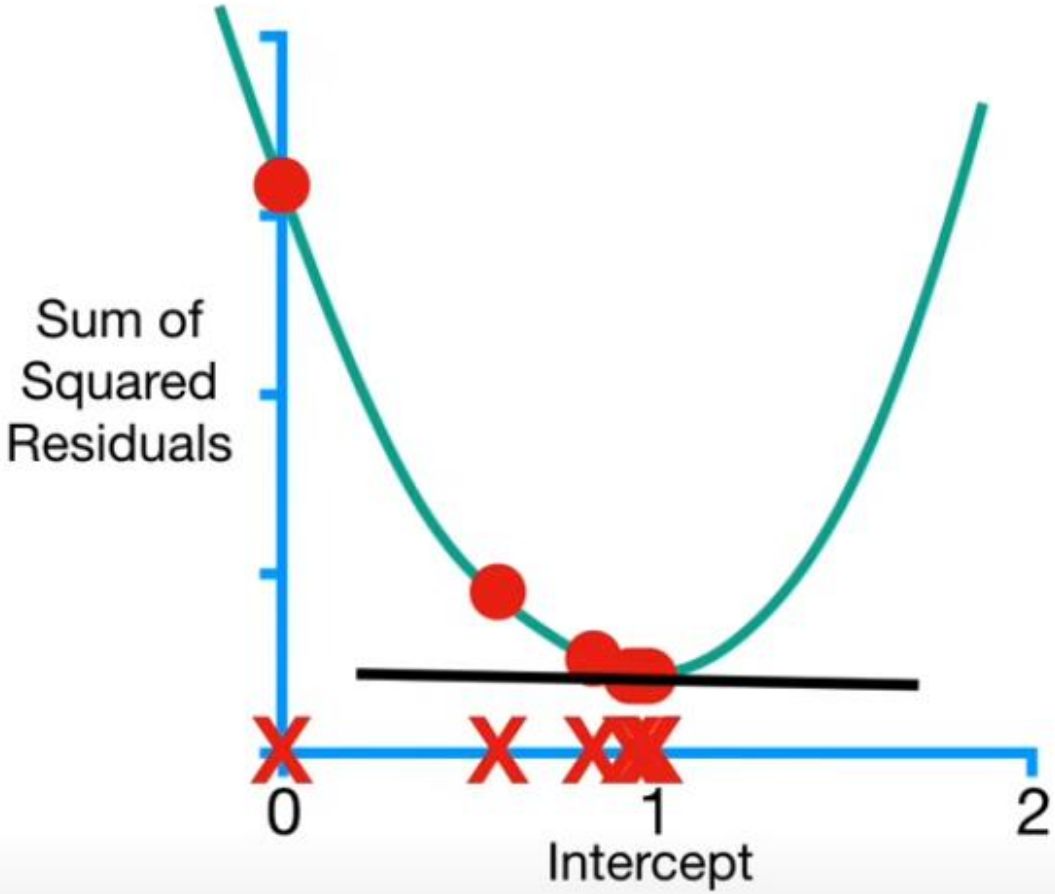
In practice, the
Minimum Step Size = 0.001
or smaller.

Step Size = Slope x Learning Rate



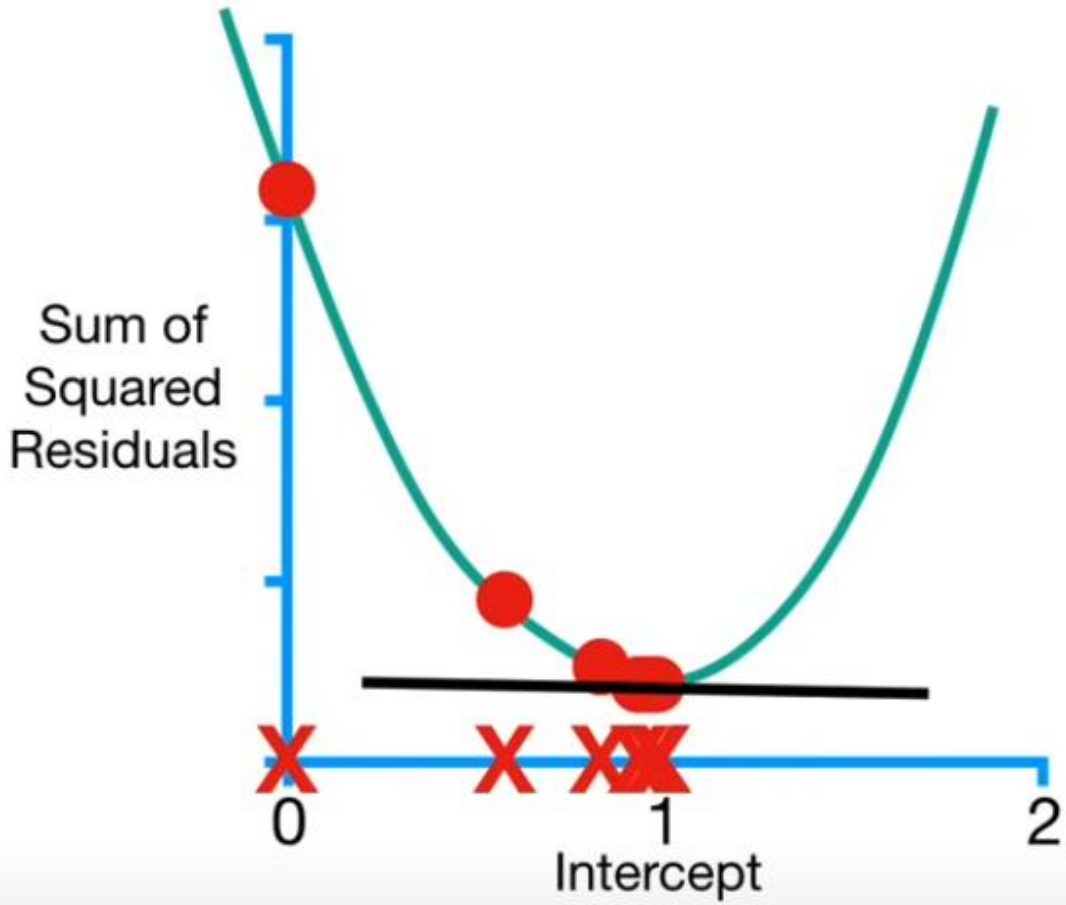
Then we would plug in
0.009 for the **Slope** and **0.1**
for the **Learning Rate**..

Step Size = **Slope** x **Learning Rate**

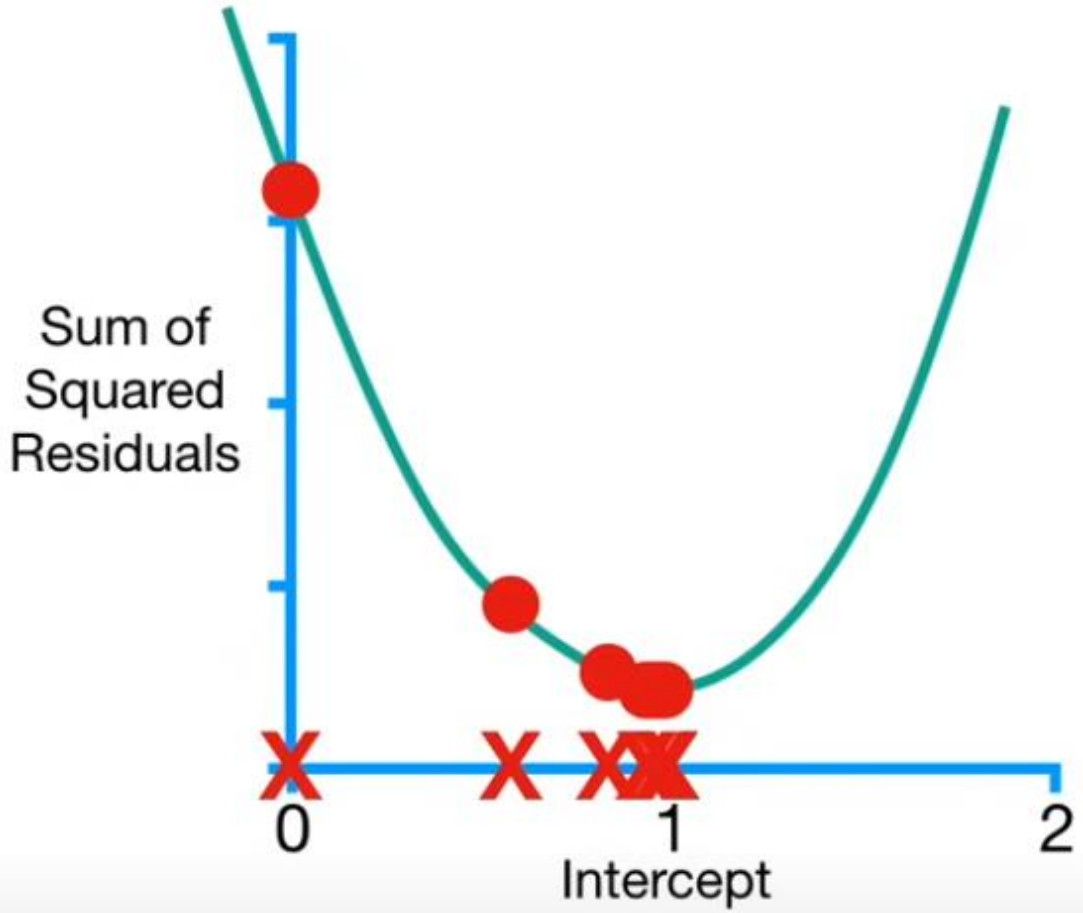


…and get **0.0009**, which is smaller than **0.001**, so **Gradient Descent** would stop.

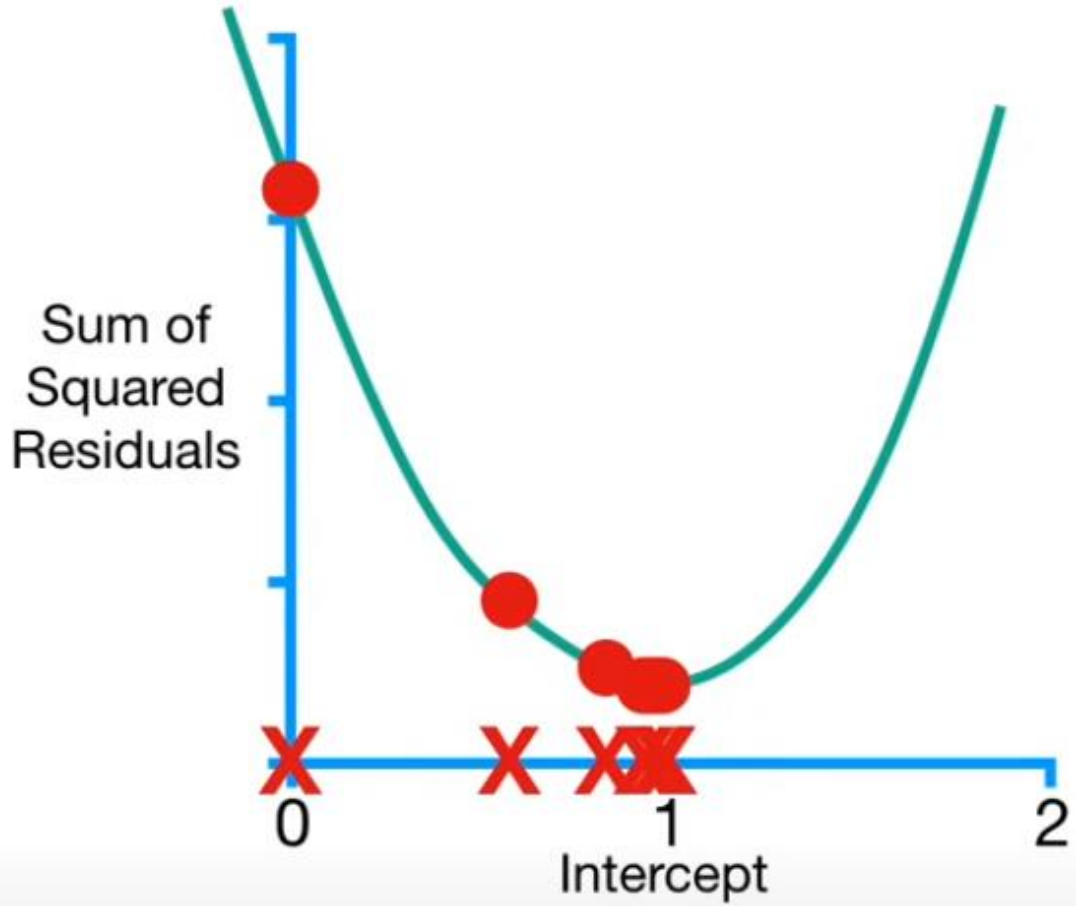
Step Size = **0.009** × **0.1** = 0.0009



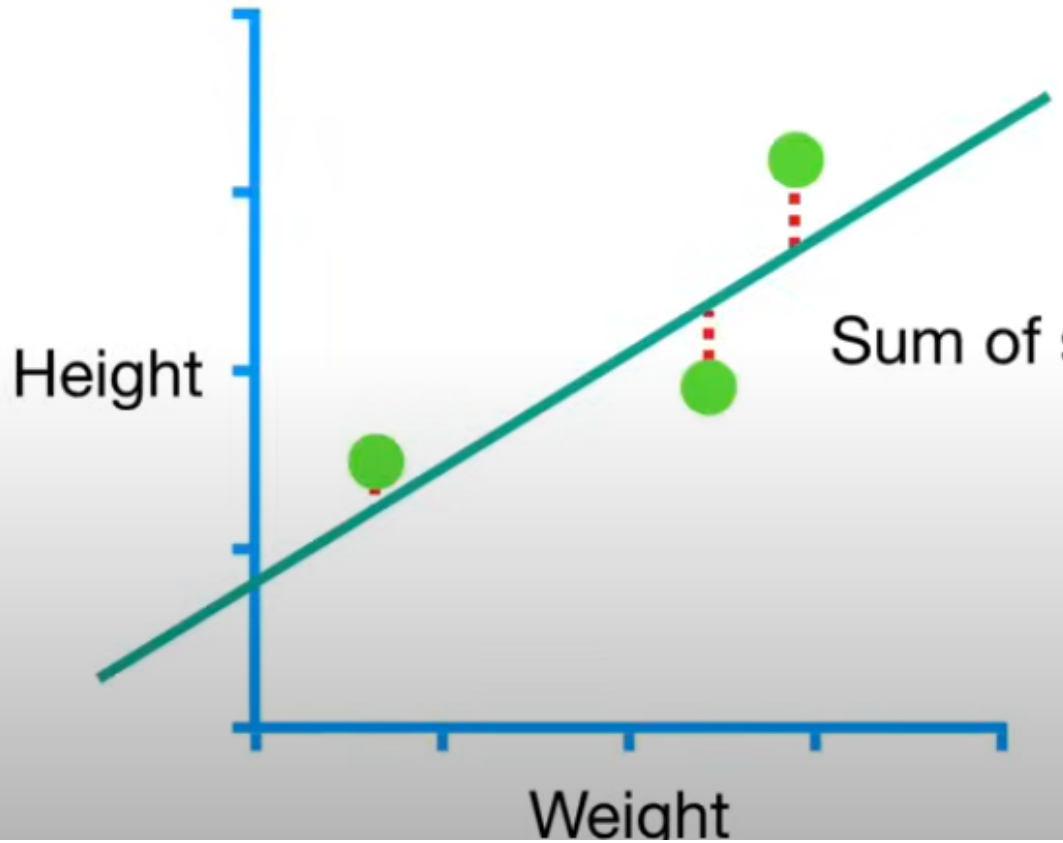
That said, **Gradient Descent** also includes a limit on the number of steps it will take before giving up.



So, even if the **Step Size** is large, if there have been more than the **Maximum Number of Steps**, **Gradient Descent** will stop.



The first thing we did is decide to use the Sum of the Squared Residuals as the **Loss Function** to evaluate how well a line fits the data…

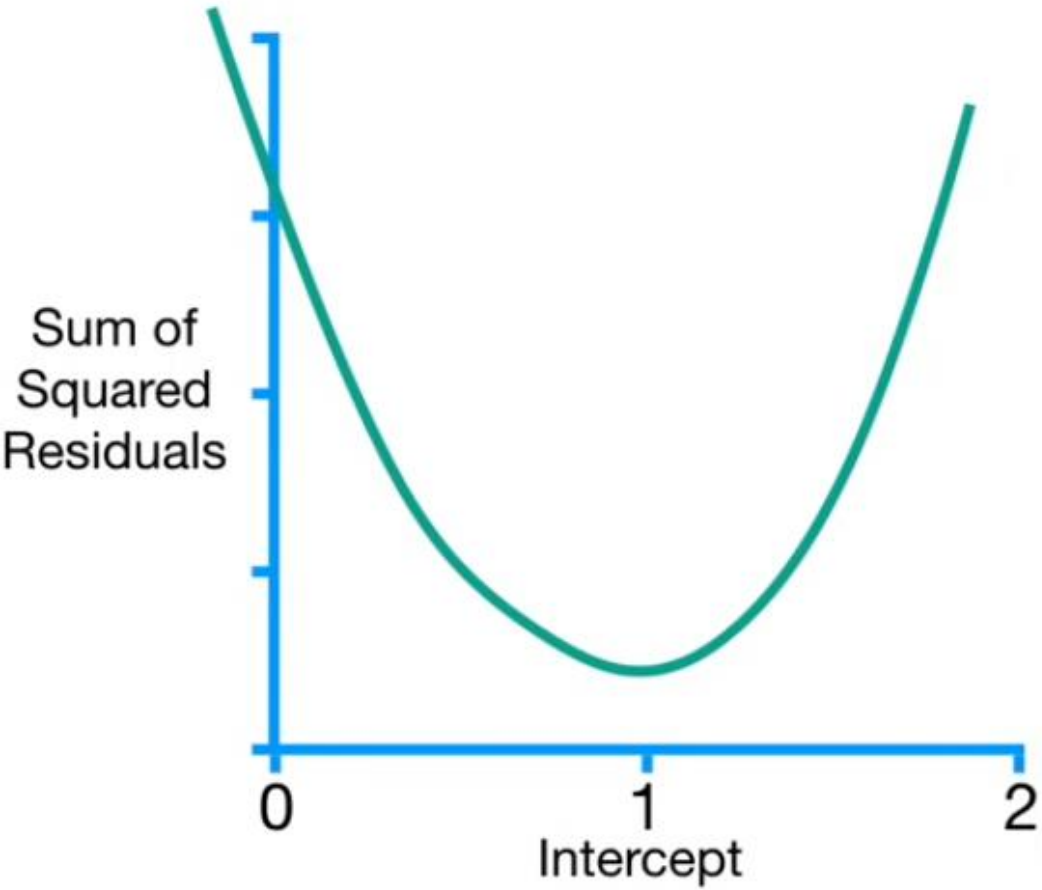


Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

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

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

…then we took the derivative of the Sum of the Squared Residuals. In other words, we took the derivative of the **Loss Function**…



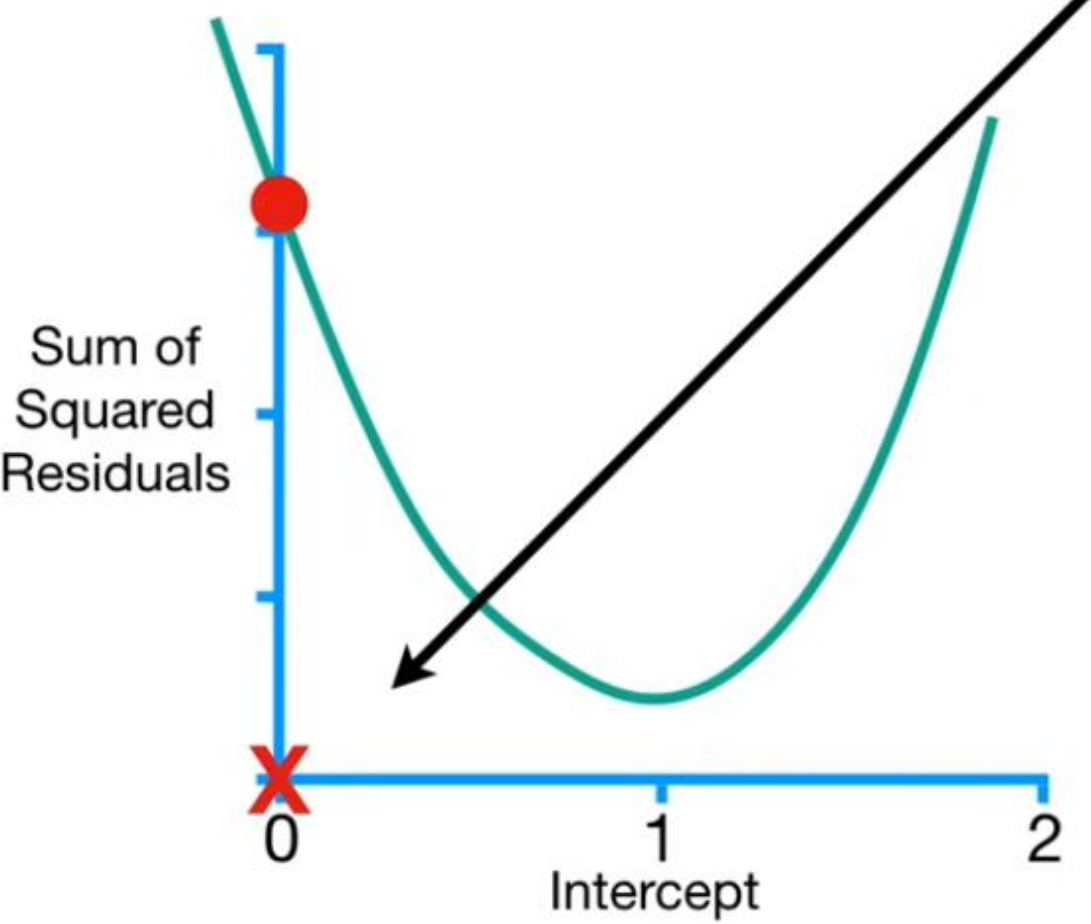
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

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

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

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

…then we picked a random value for the **Intercept**, in this case we set the **Intercept** = **0**…



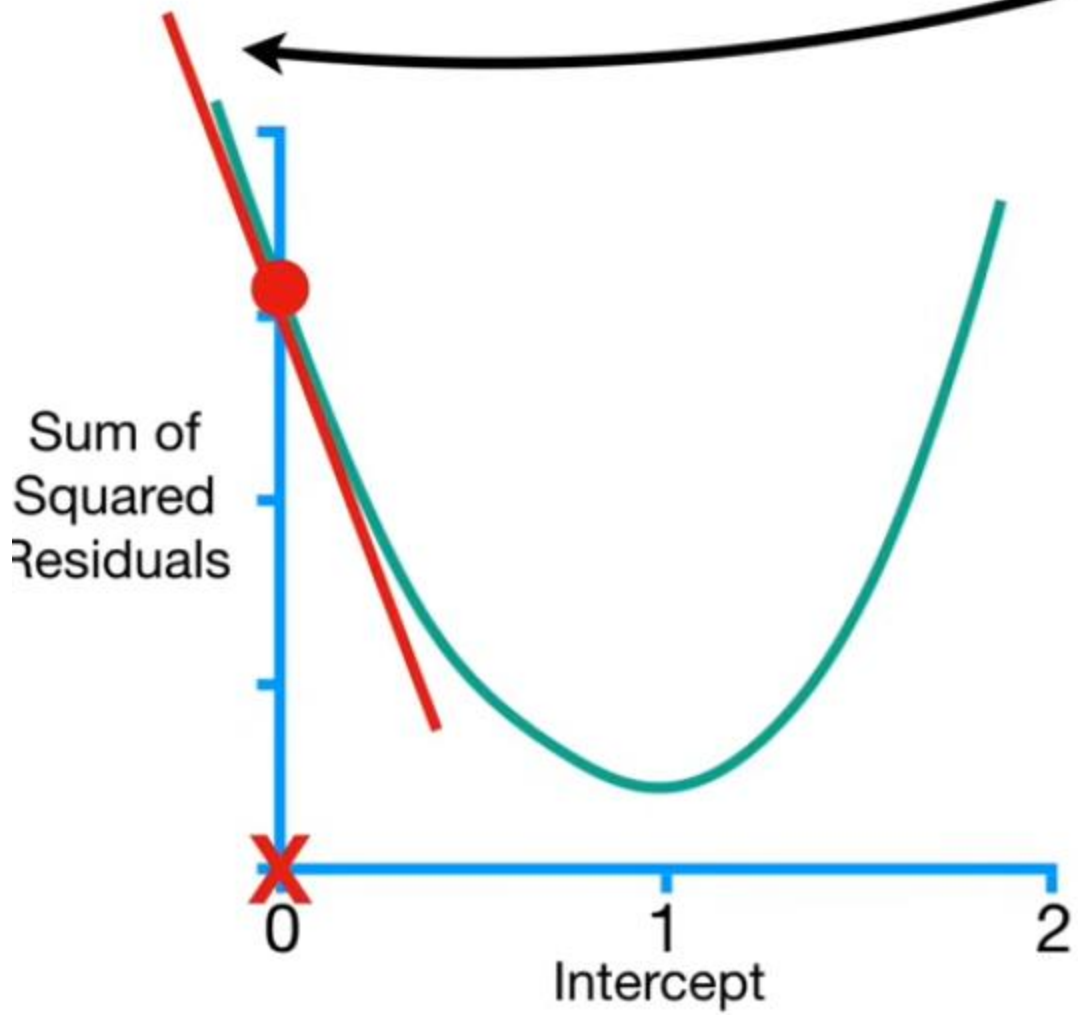
$$\frac{d}{d\ intercept}\ \text{Sum of squared residuals} =$$

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

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

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

…then we calculated the derivative
when the **Intercept** = **0**…



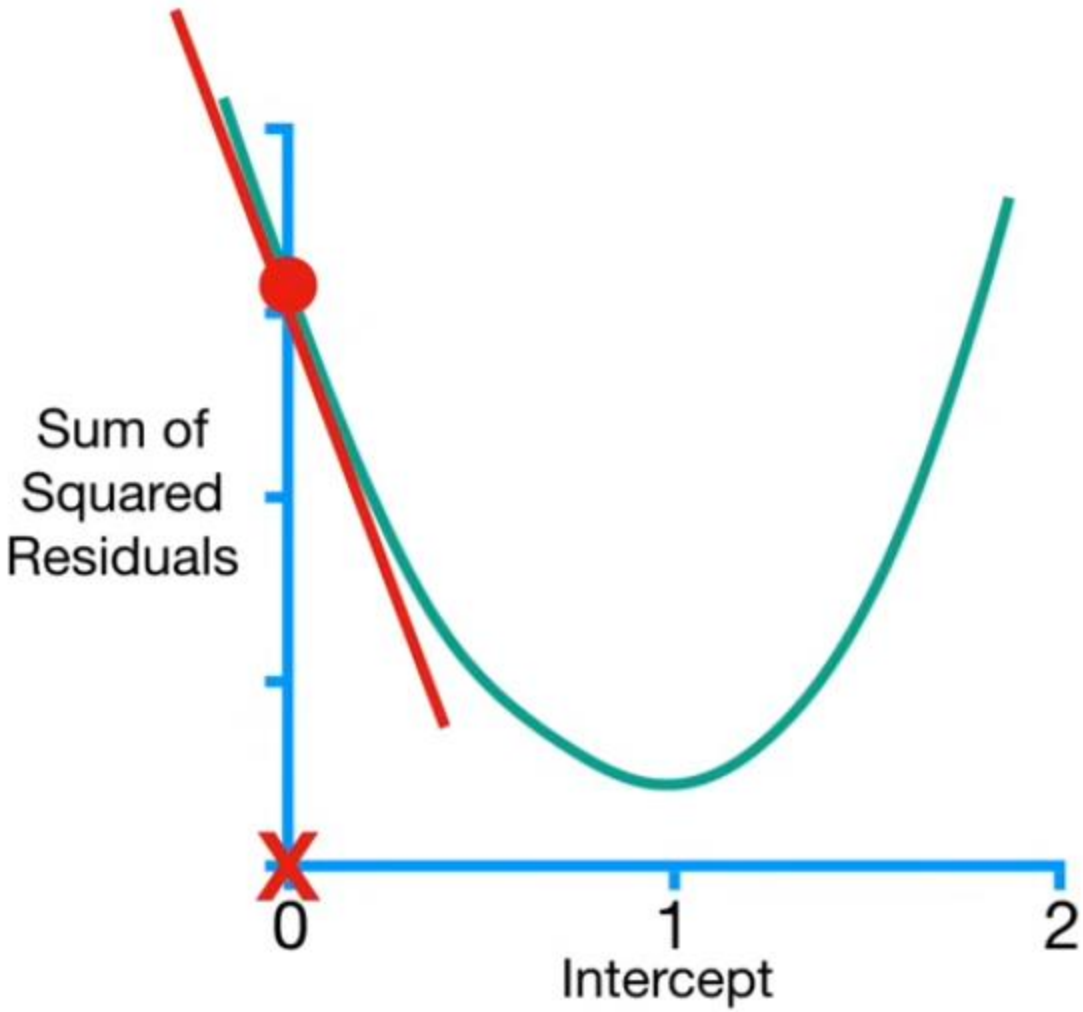
$$\frac{d}{d\ intercept}\ \text{Sum of squared residuals} =$$

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

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

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

…plugged that slope into the **Step Size** calculation…

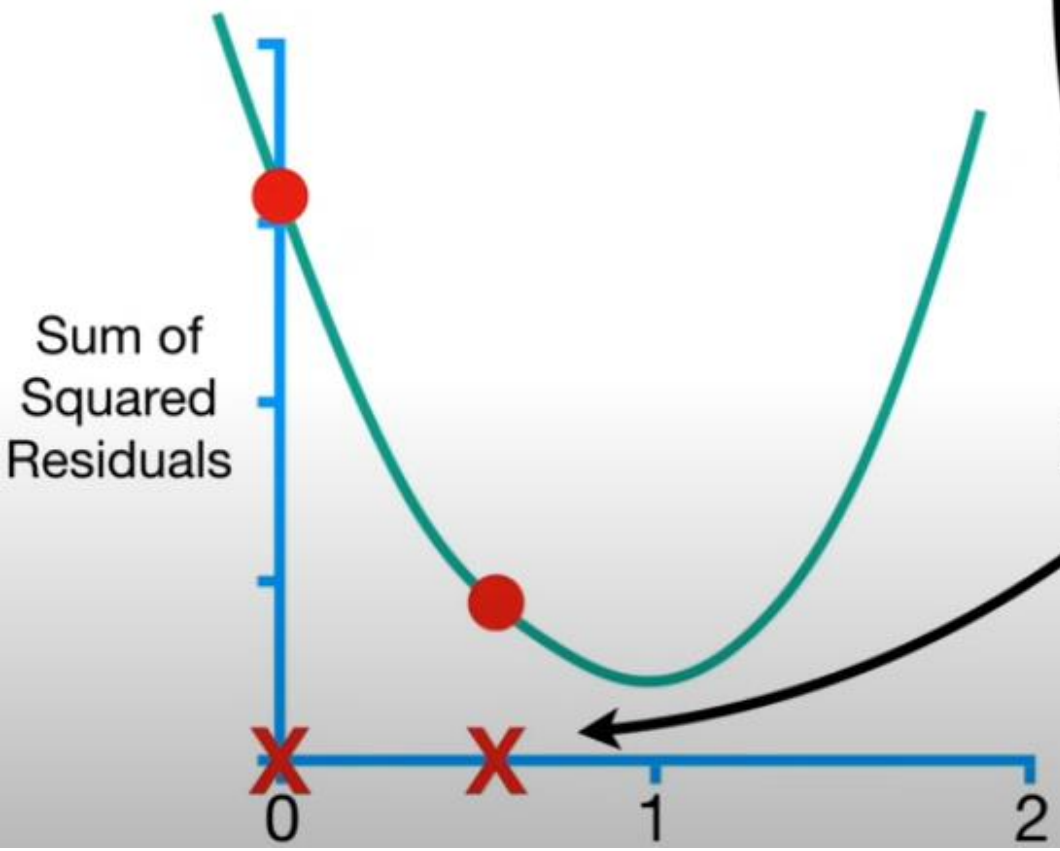


Step Size = **Slope** x **Learning Rate**

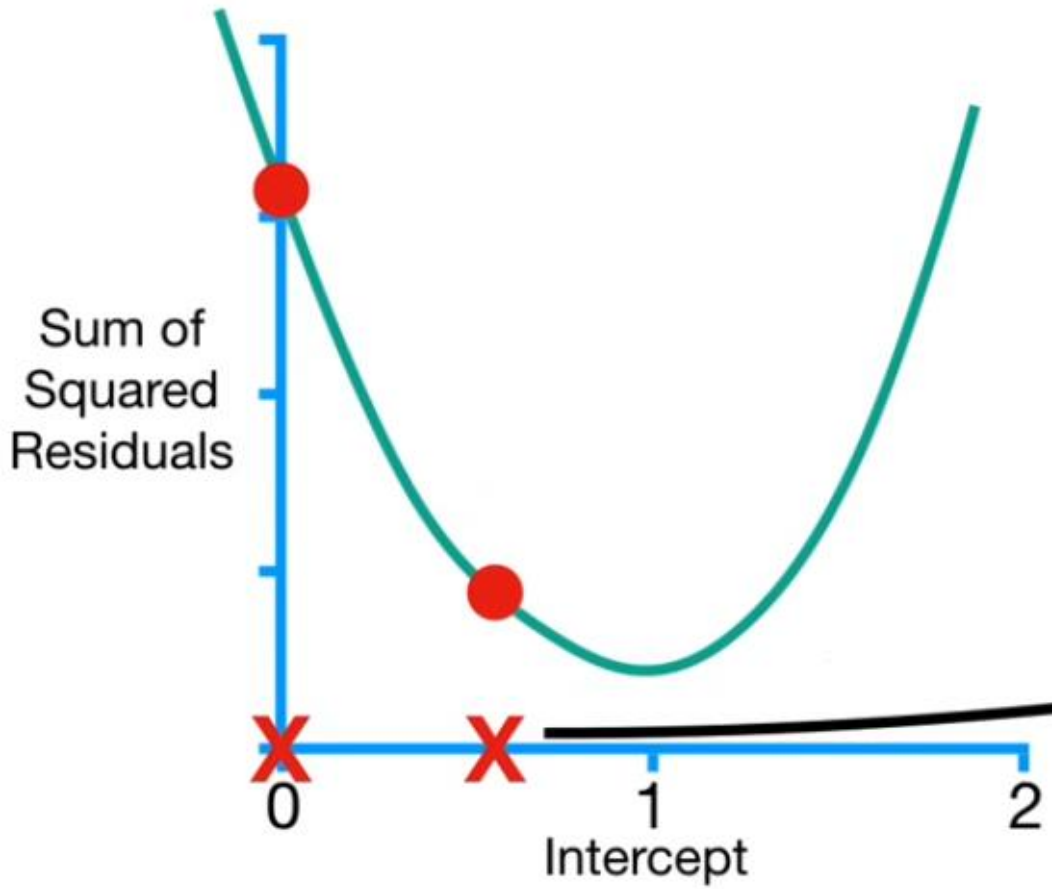
…then calculated the **New Intercept**, the difference between the **Old Intercept** and the **Step Size**.

Step Size = Slope × Learning Rate

New Intercept = Old Intercept - Step Size



Lastly, we plugged the **New Intercept** into the derivative and repeated everything until **Step Size** was close to **0**.



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

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

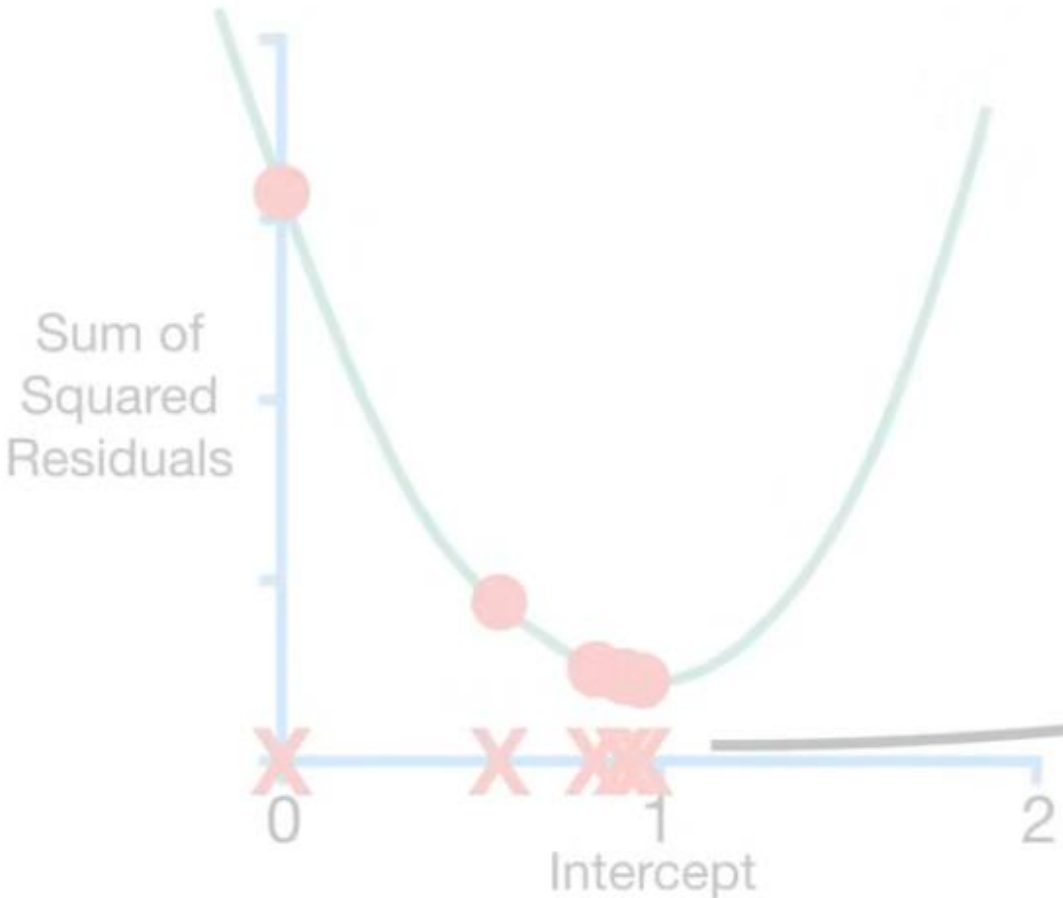
$$+ -2(\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))$$

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

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

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

DOUBLE BAM!!!



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

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

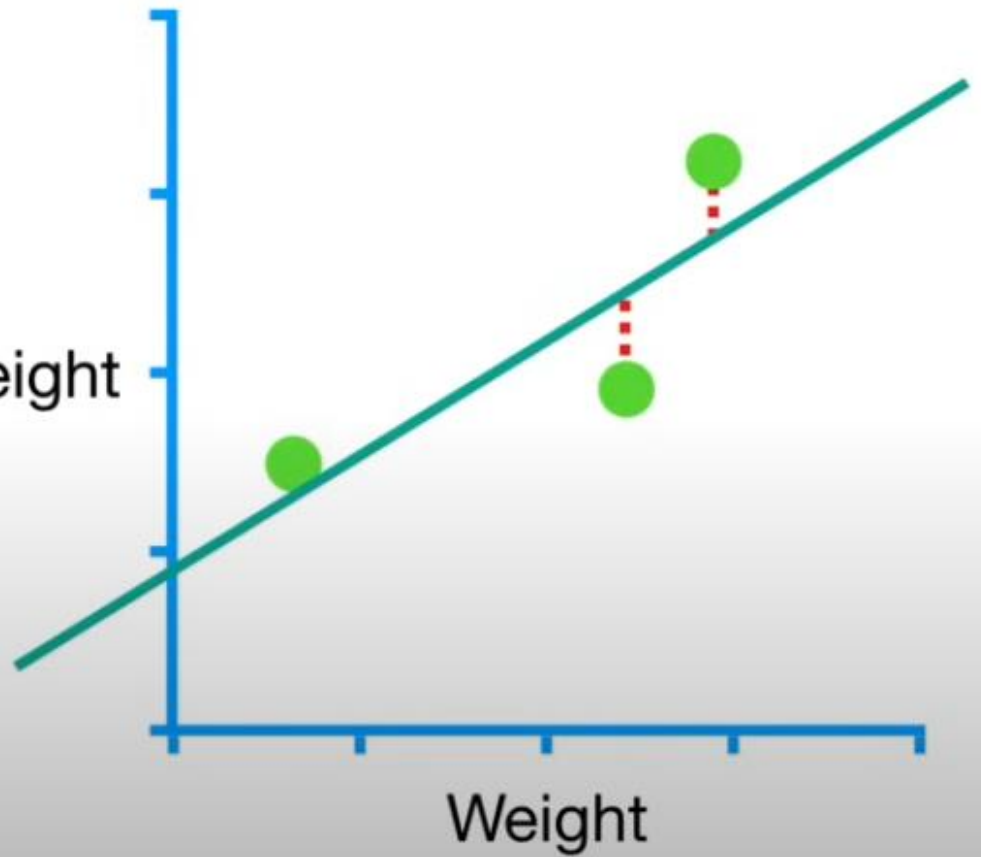
$$+ -2(\mathbf{1.9} - (\text{intercept} + 0.64 \times \mathbf{2.3}))$$

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

Step Size = Slope × Learning Rate

New Intercept = Old Intercept - Step Size

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

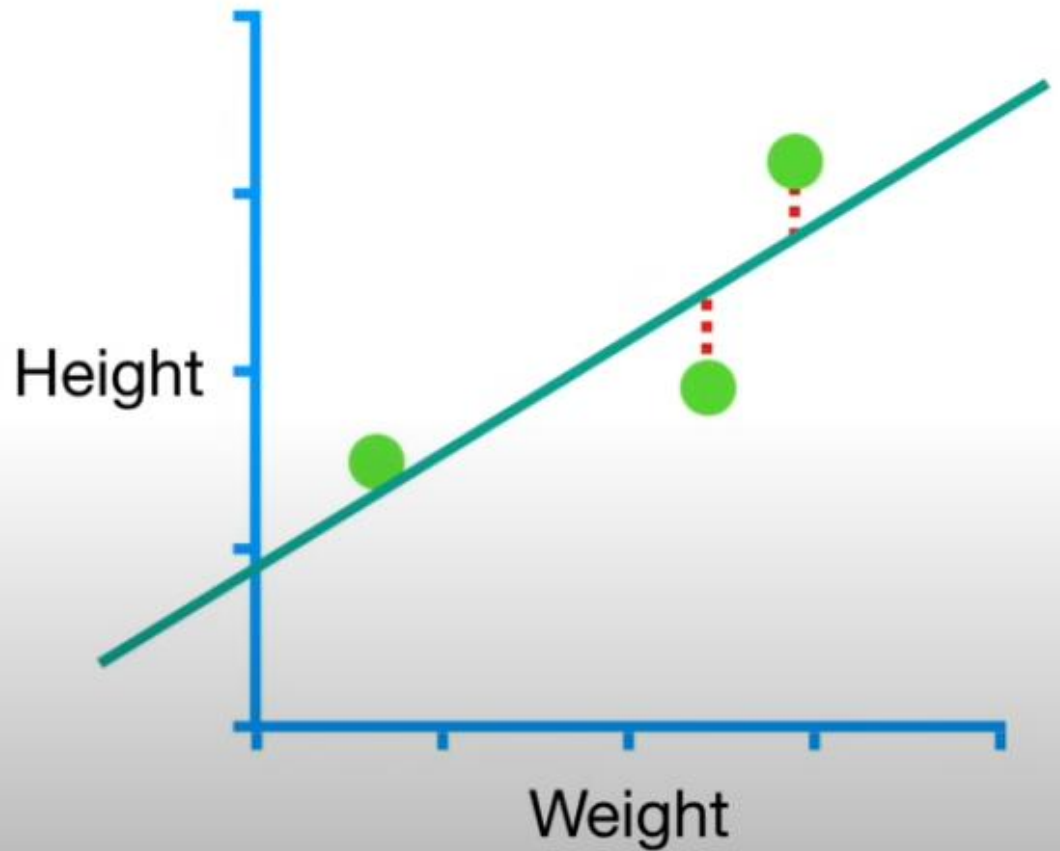


…let's talk abut how to estimate the **Intercept** *and* the **Slope**.

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

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

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



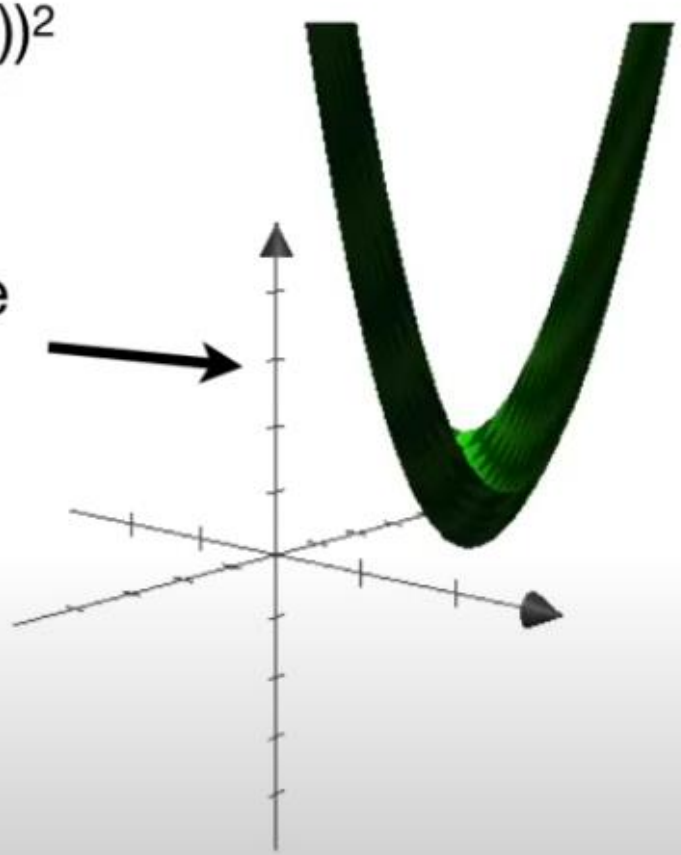
Just like before, we will use the Sum of the Squared Residuals as the **Loss Function**

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

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

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

This axis is the Sum of the Squared Residuals…

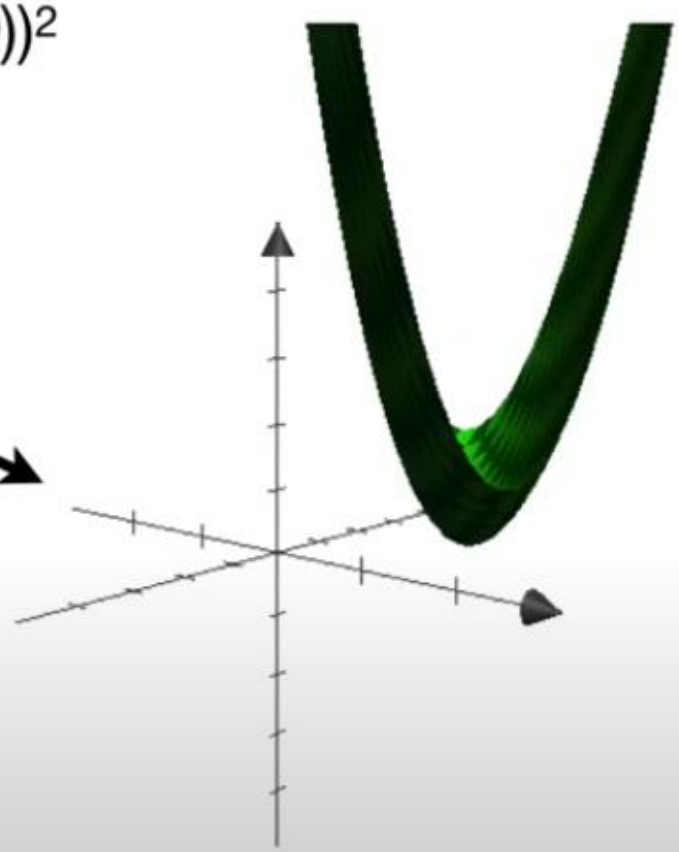


Sum of squared residuals = (**1.4** - (intercept + slope × **0.5**))^2

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

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

…this axis represents different values for the **Slope**…

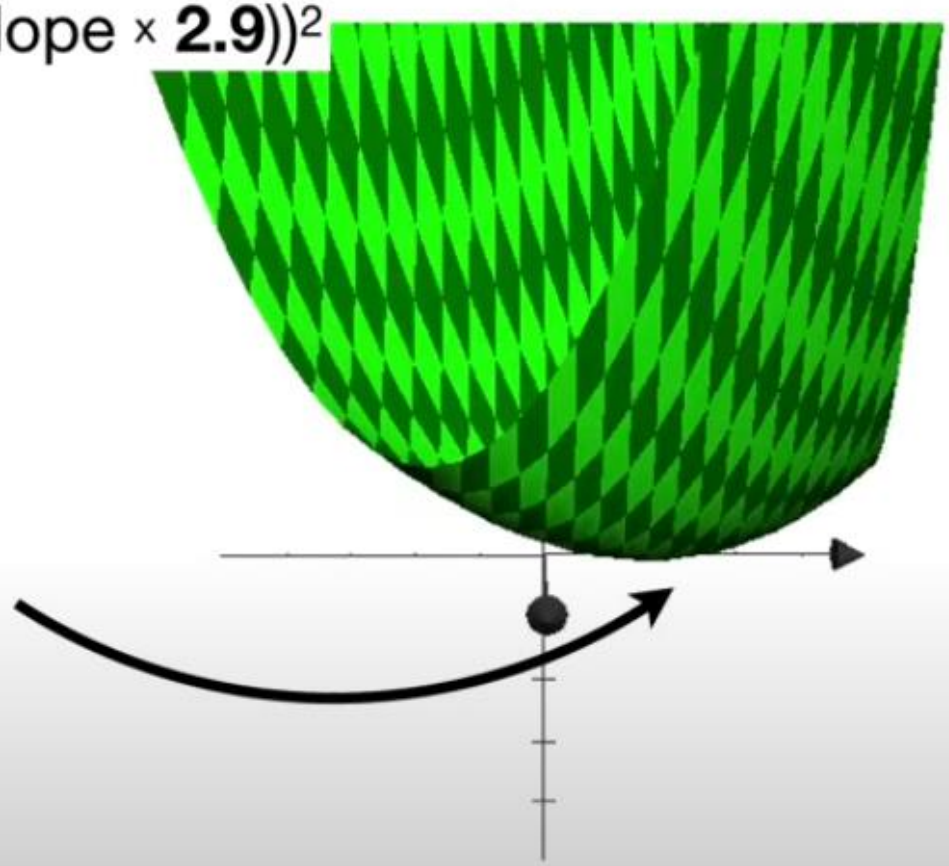


Sum of squared residuals = (**1.4** - (intercept + slope × **0.5**))^2

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

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

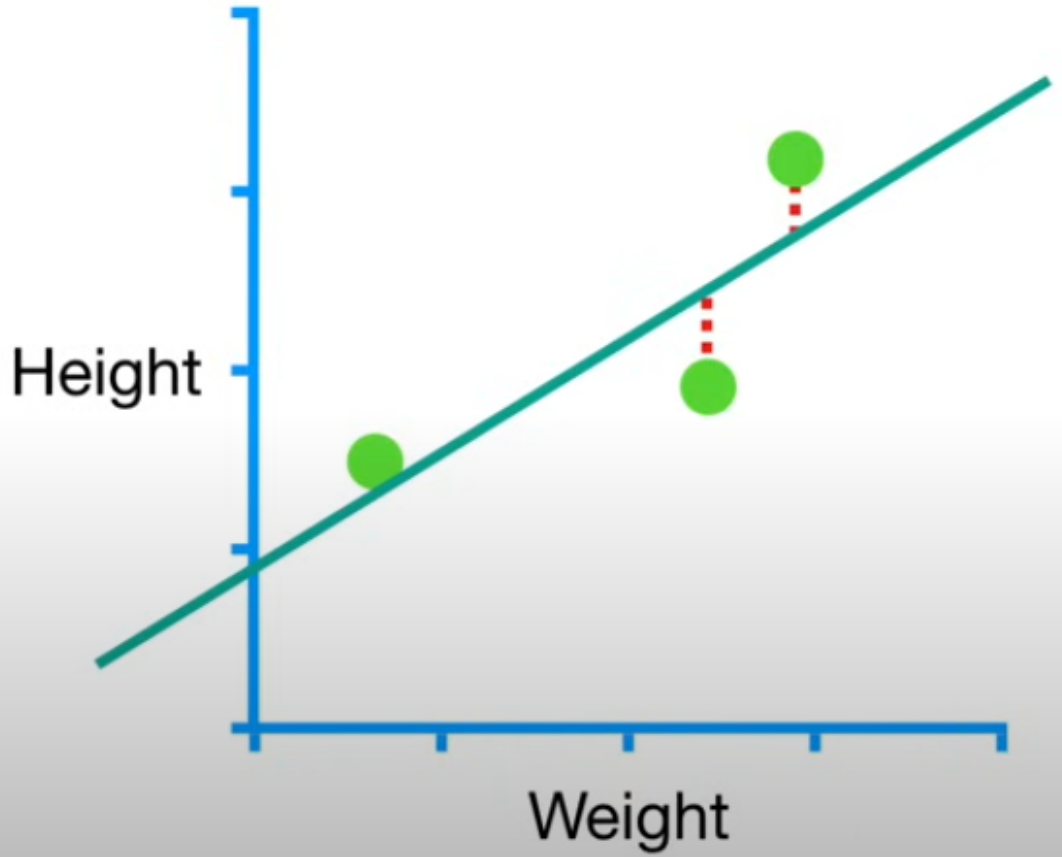
We want to find the values for the **Intercept** and **Slope** that give us the minimum Sum of the Squared Residuals.



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

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

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

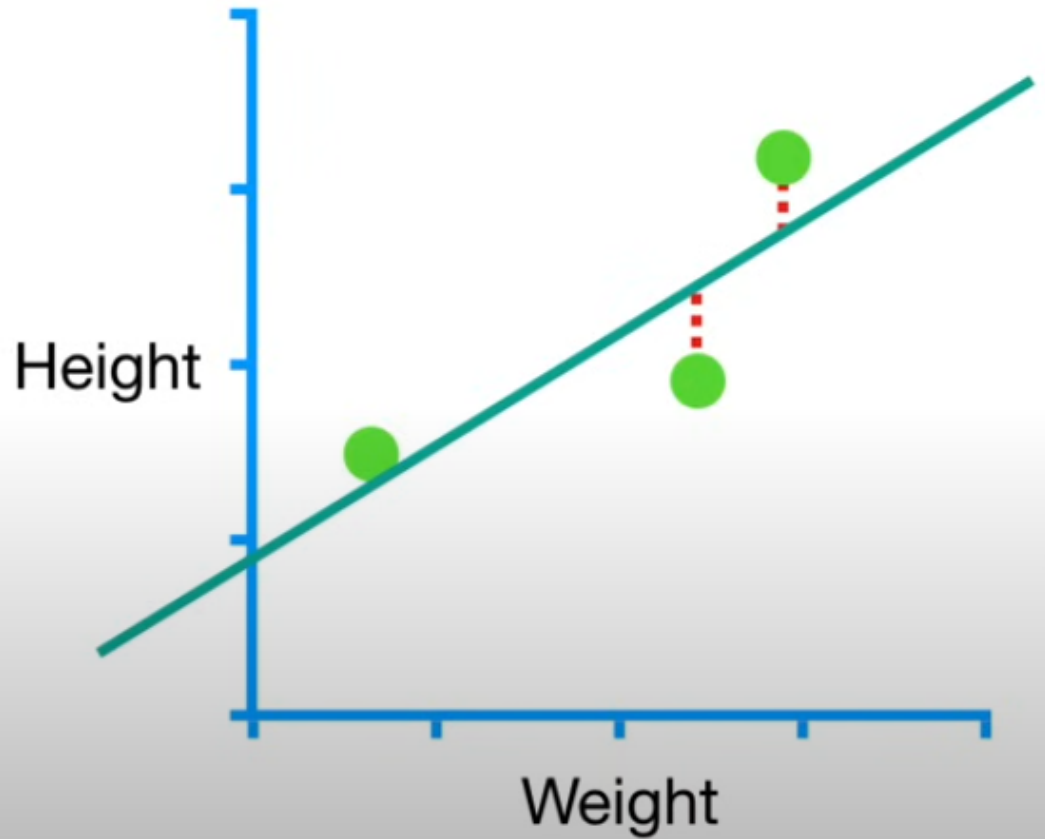


So, just like before, we need to take the derivative of this function…

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

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

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



…and just like before, we'll take the derivative with respect to the **Intercept**…

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

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

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

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

We'll start by taking the derivative with respect to the intercept.

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

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

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

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

Just like before, we take the derivative of each part…

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

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

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

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

…and move the square to the front…

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

…and this…

…is the derivative of the first part…

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

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

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

Likewise, we replace these terms with their derivatives…

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

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

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

…and this whole thing is the derivative of the Sum of the Squared Residuals with respect to the **Intercept**.

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

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

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

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

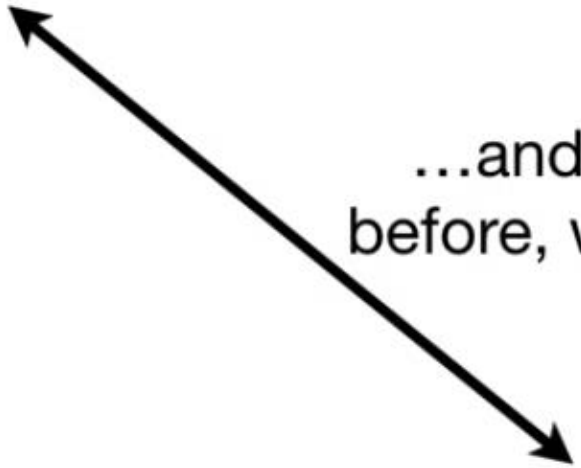
+ (**1.9** - (intercept + slope × **2.3**))^2

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

Now let's take the derivative of the Sum of the Squared Residuals with respect to the **Slope**.

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

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



…and just like
before, we'll use…

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

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

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

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

…and multiply that by the derivative of the stuff inside the parentheses.

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

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

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

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**.

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

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

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

So we end up with **-0.5**.

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

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

…then we simplify by moving the -0.5 to the front

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

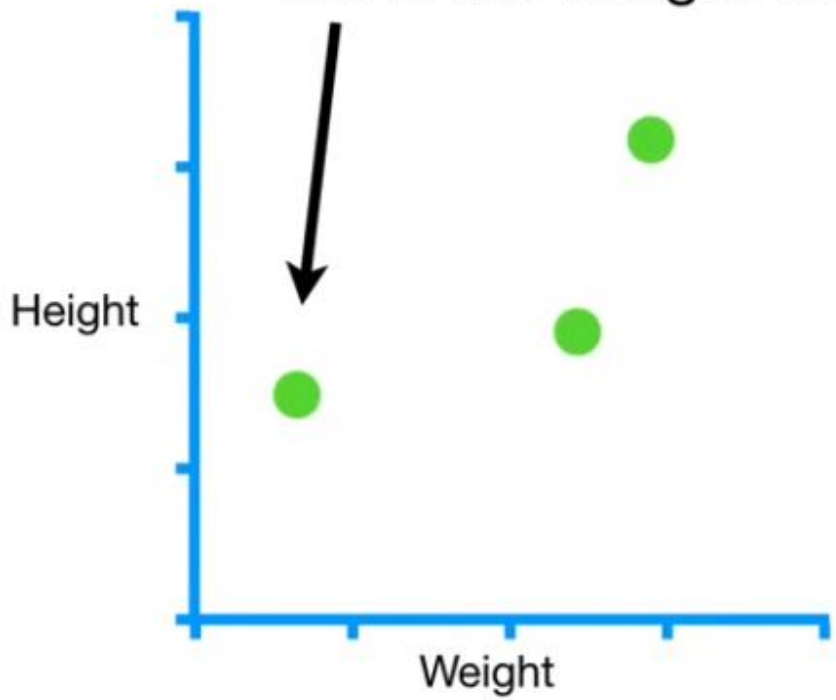
$$+ \frac{d}{d\,slope}\,(\mathbf{1.9} - (\text{intercept} + \text{slope} \times \mathbf{2.3}))^2$$

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

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

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

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.



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

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

…and this…

…is the derivative of the first part…

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

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

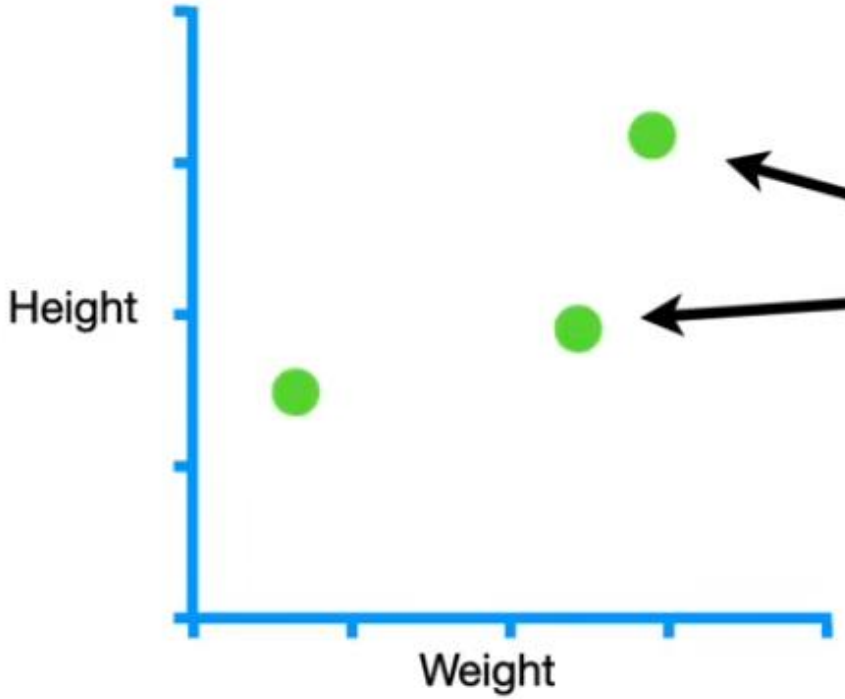
$$+ \frac{d}{d\,slope}\,(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$

Likewise, we replace these terms with their derivatives.

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

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

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



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

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

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

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

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

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

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

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

Here's the derivative of the Sum of the Squared Residuals with respect to the **Intercept**...

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

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

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

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

NOTE: When you have two or more derivatives of the same function, they are called a **Gradient**.

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

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

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

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

$$\frac{d}{d\ intercept}$$ 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…

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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**.

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

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

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

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

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

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

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

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

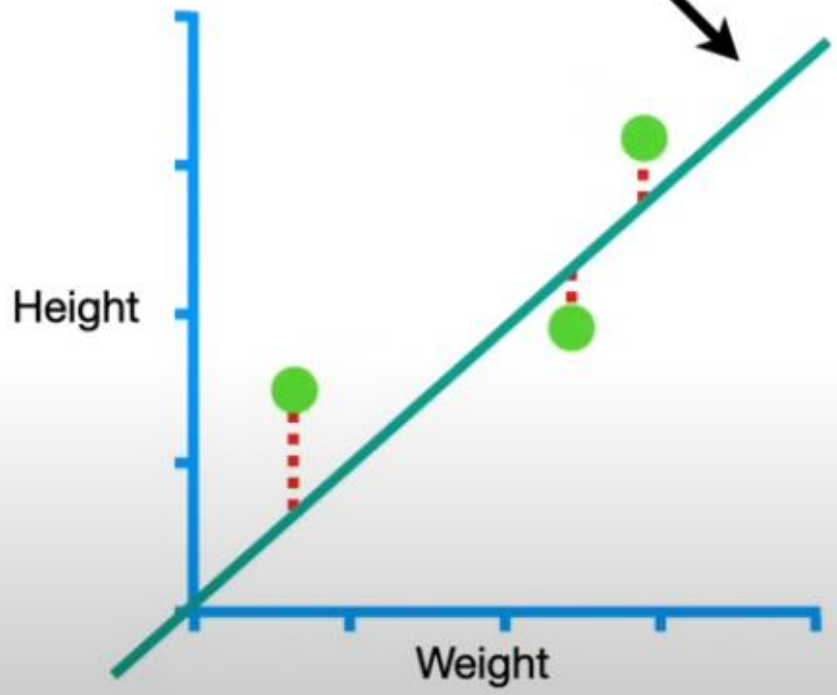
$$\frac{d}{d\ slope}$$ Sum of squared residuals =

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

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

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

Thus, this line, with **Intercept = 0** and **Slope = 1**, is where we will start.



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

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

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

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

…and that gives us two **Slopes**…

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

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

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

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

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

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

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

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

$$\mathbf{Step\ Size}_{Intercept} = -1.6 \times \text{Learning Rate}$$

…and multiply by the **Learning Rate**, which this time we set to **0.01**…

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

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

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

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

$$\mathbf{Step\ Size}_{Slope} = -0.8 \times \text{Learning Rate}$$

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

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

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

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

$$\textbf{Step Size}_{\text{Intercept}} = -1.6 \times 0.01$$

NOTE: The larger **Learning Rate** that we used in the first example doesn't work this time. Even after a bunch of steps, **Gradient Descent** doesn't arrive at the correct answer.

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

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

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

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

$$\textbf{Step Size}_{\text{Slope}} = -0.8 \times 0.01$$

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

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

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

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

$$\textbf{Step Size}_{\text{Intercept}} = -1.6 \times 0.01$$

The good news is that in practice, a reasonable **Learning Rate** can be determined automatically by starting large and getting smaller with each step.

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

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

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

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

$$\textbf{Step Size}_{\text{Slope}} = -0.8 \times 0.01$$

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

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

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

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

$$\mathbf{Step\ Size}_{Intercept} = -1.6 \times 0.01 = \mathbf{-0.016}$$

Anyway, we do the math and get two **Step Sizes**.

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

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

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

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

$$\mathbf{Step\ Size}_{Slope} = -0.8 \times 0.01 = \mathbf{-0.008}$$

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

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

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

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

$\mathbf{Step\ Size}_{Intercept} = -1.6 \times 0.01 = \mathbf{-0.016}$

New Intercept = 0 - **Step Size**

Now we calculate the **New Intercept** and **New Slope** by plugging in the **Old Intercept** and the **Old Slope**…

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

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

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

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

$\mathbf{Step\ Size}_{Slope} = -0.8 \times 0.01 = \mathbf{-0.008}$

New Slope = 1 - **Step Size**

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

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

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

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

$\mathbf{Step\ Size}_{Intercept} = -1.6 \times 0.01 = \mathbf{-0.016}$

New Intercept = 0 - (-0.016) = 0.016

…and we end up with a **New Intercept** and a **New Slope**.

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

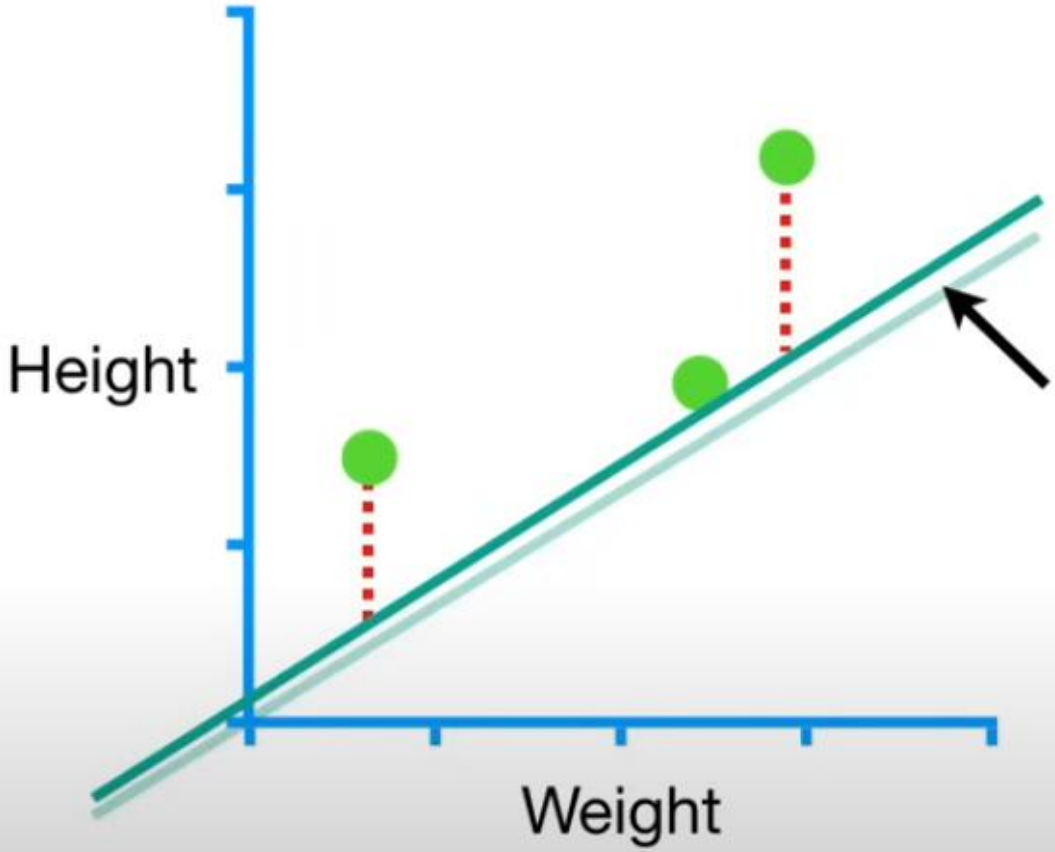
$-2 \times \mathbf{0.5}(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5}))$

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

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

$\mathbf{Step\ Size}_{Slope} = -0.8 \times 0.01 = \mathbf{-0.008}$

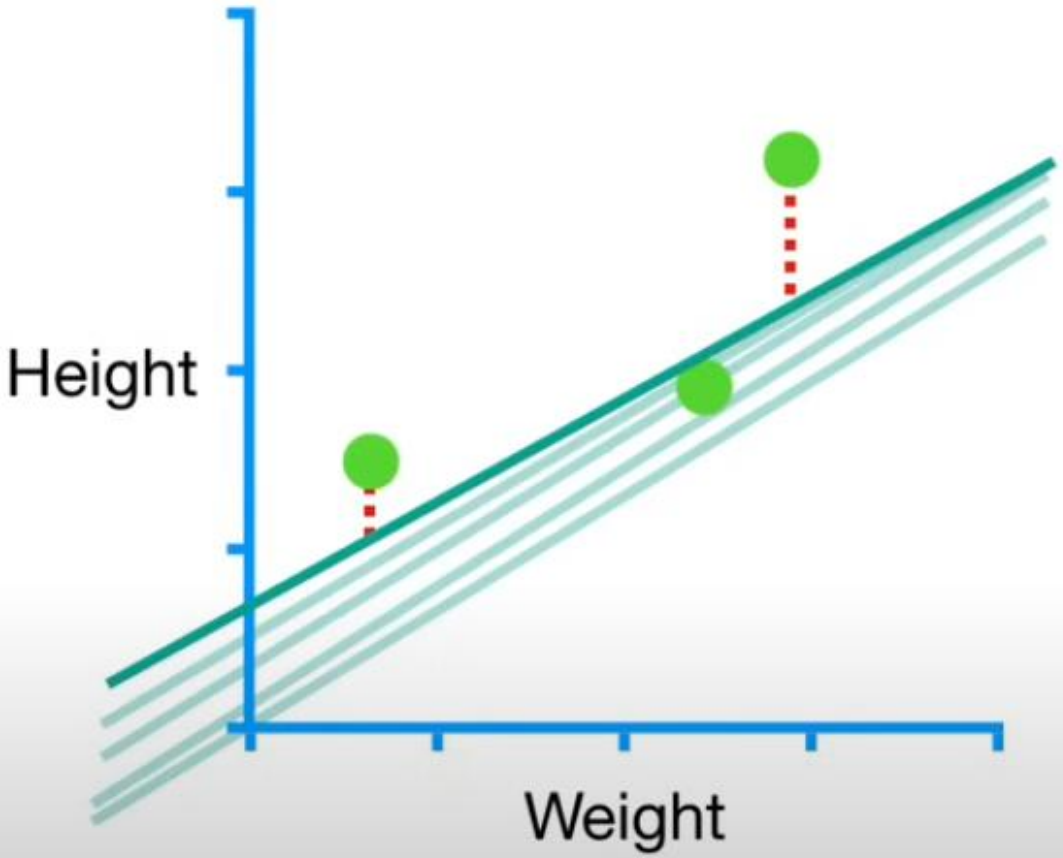
New Slope = 1 - (-0.008) = 1.008



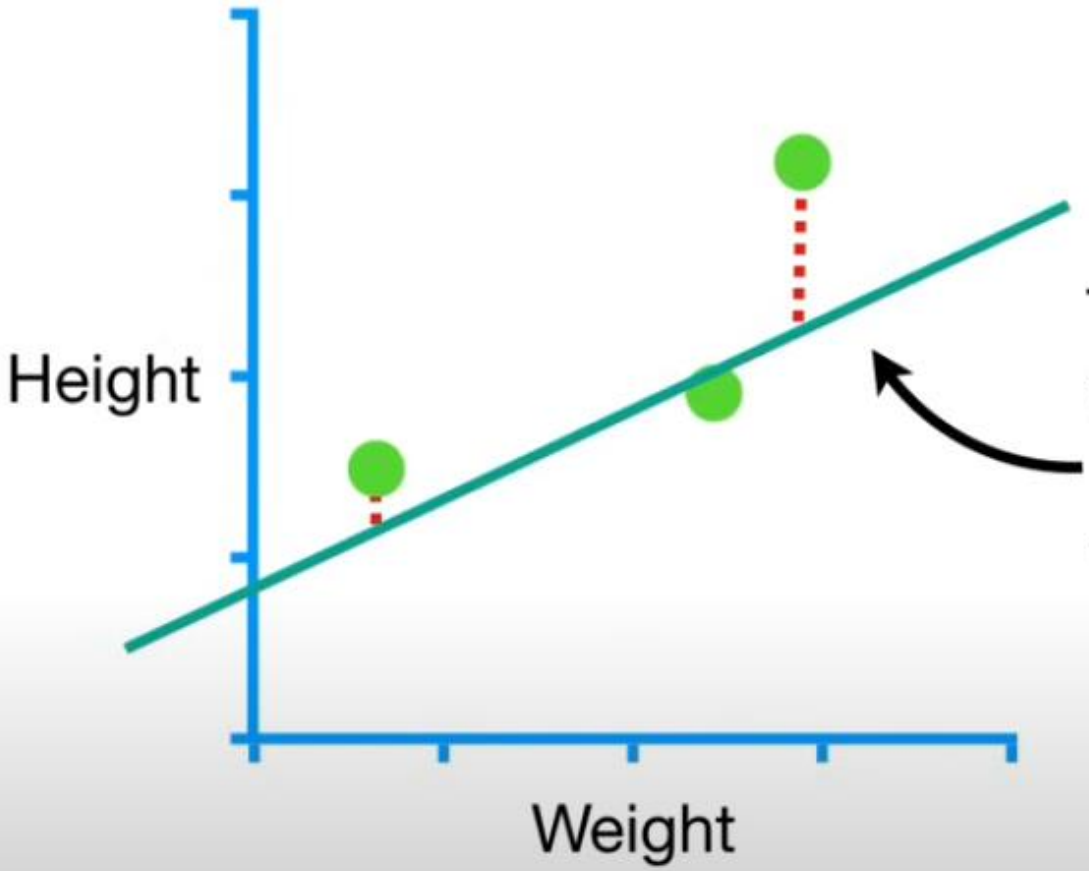
New Intercept = 0 - (-0.016) = 0.016

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

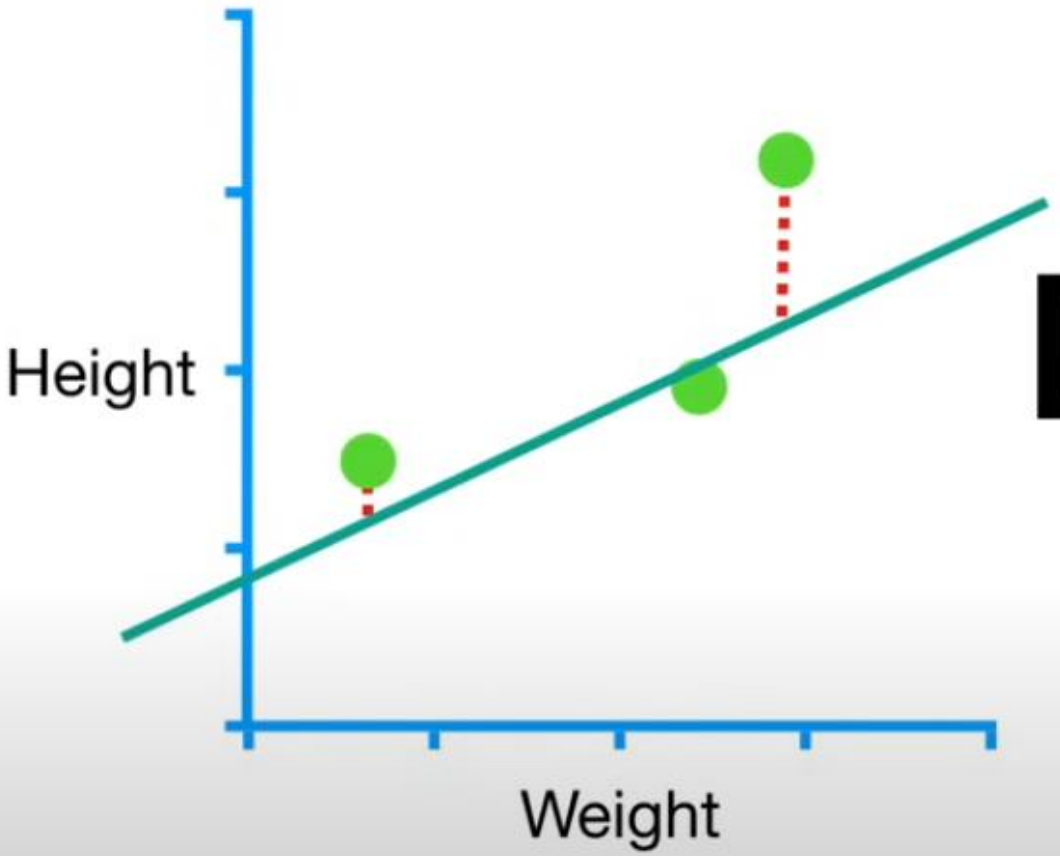
New Slope = 1 - (-0.008) = 1.008



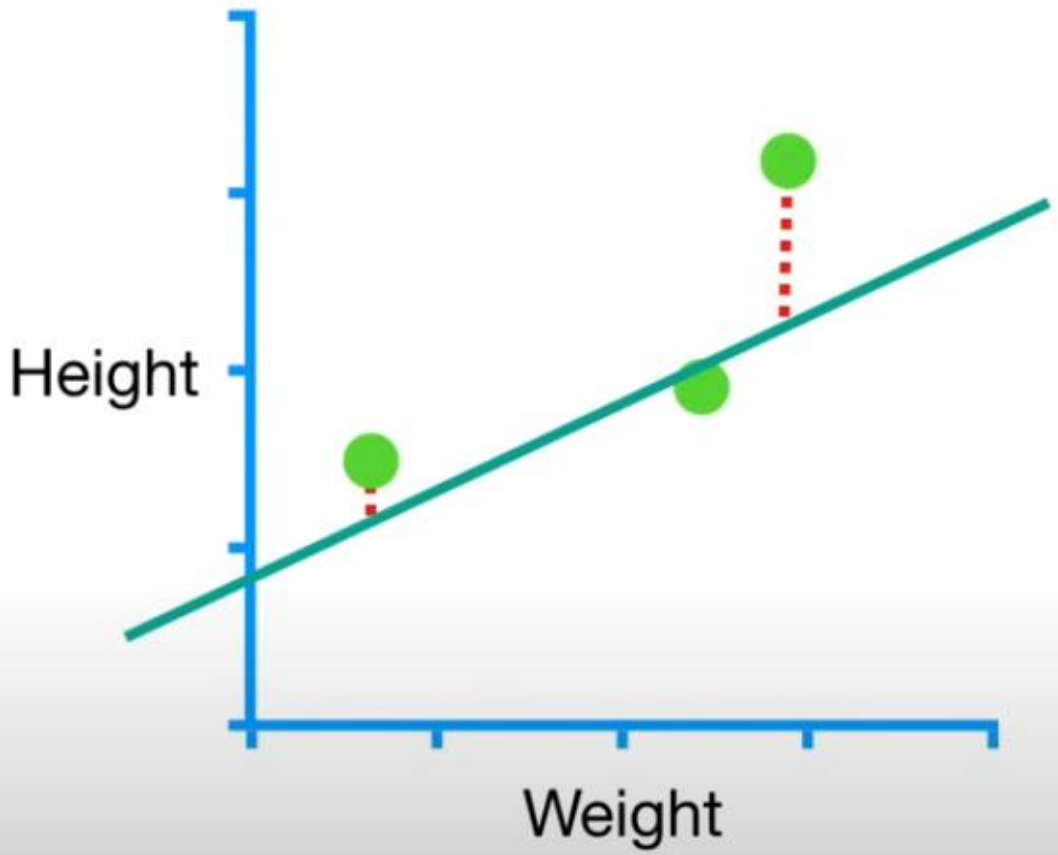
Now we just repeat what we did until all of the **Steps Sizes** are very small or we reach the **Maximum Number of Steps**.



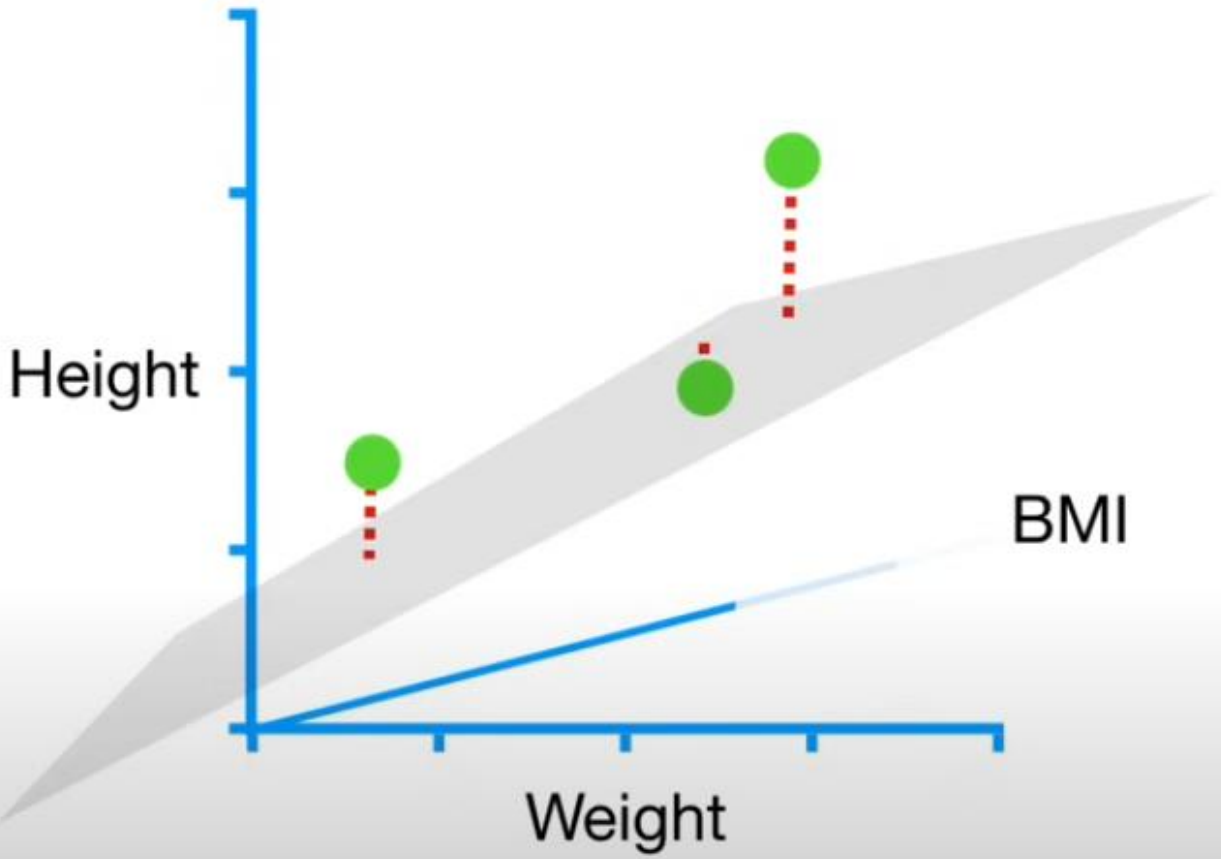
This is the best fitting line, with **Intercept** = **0.95** and **Slope** = **0.64**, the same values we get from **Least Squares**.



DOUBLE BAM!!!



We now know how **Gradient Descent** optimizes two parameters, the **Slope** and **Intercept**.



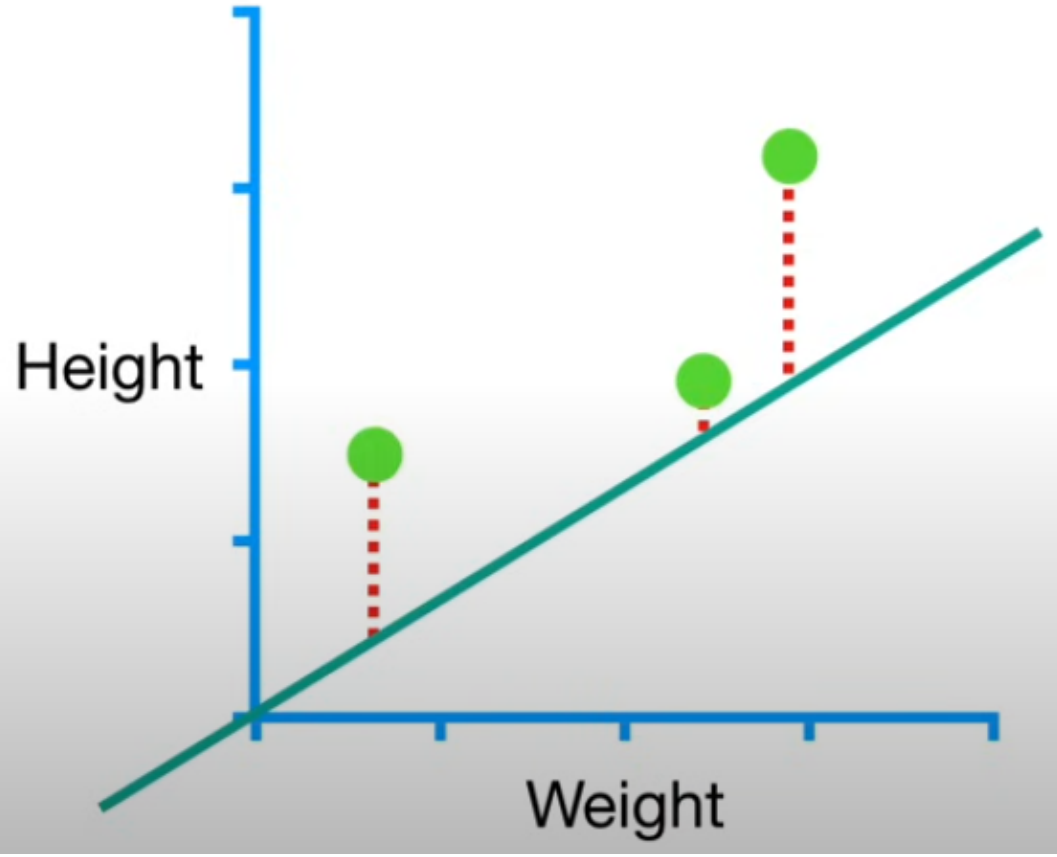
If we had more parameters, then we'd just take more derivatives and everything else stays the same.

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

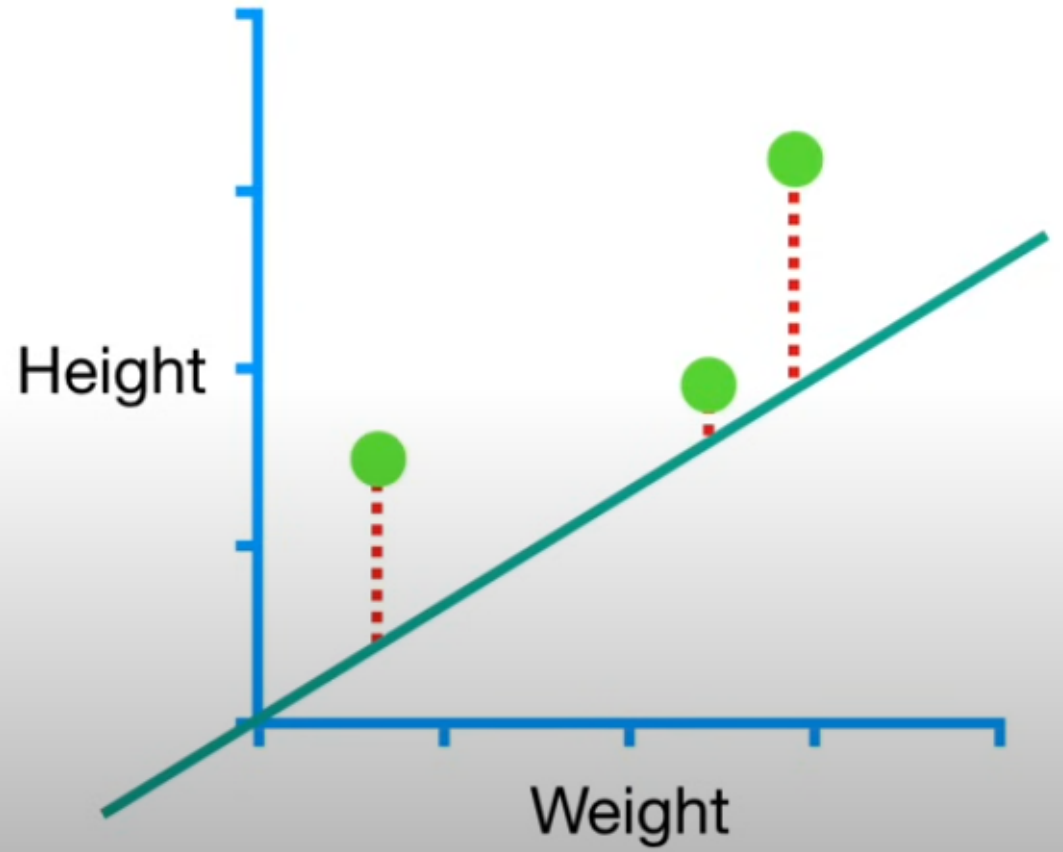
NOTE: The Sum of the Squared Residuals is just one type of **Loss Function**.



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

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 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**

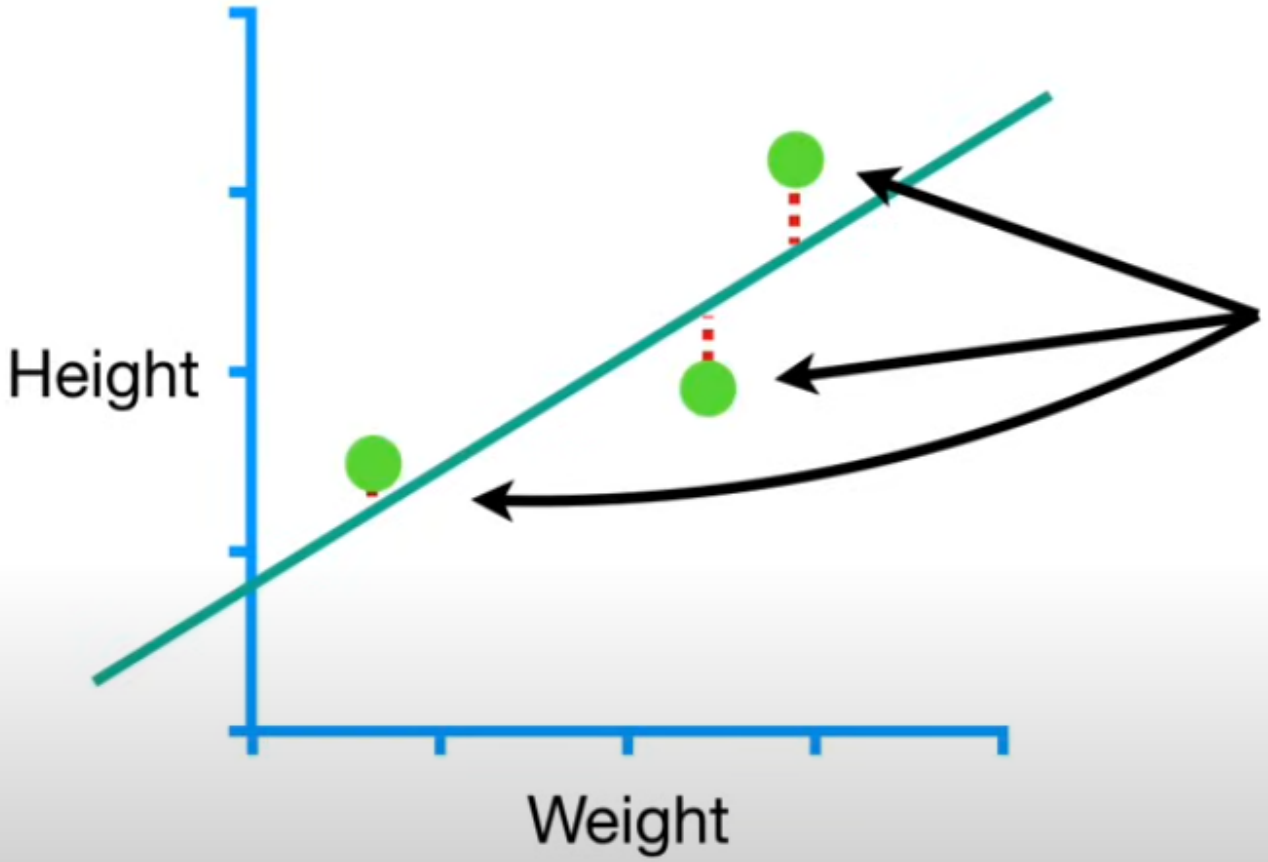
Now go back to **Step 3** and repeat until
Step Size is very small, or you reach
the **Maximum Number of Steps**.

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

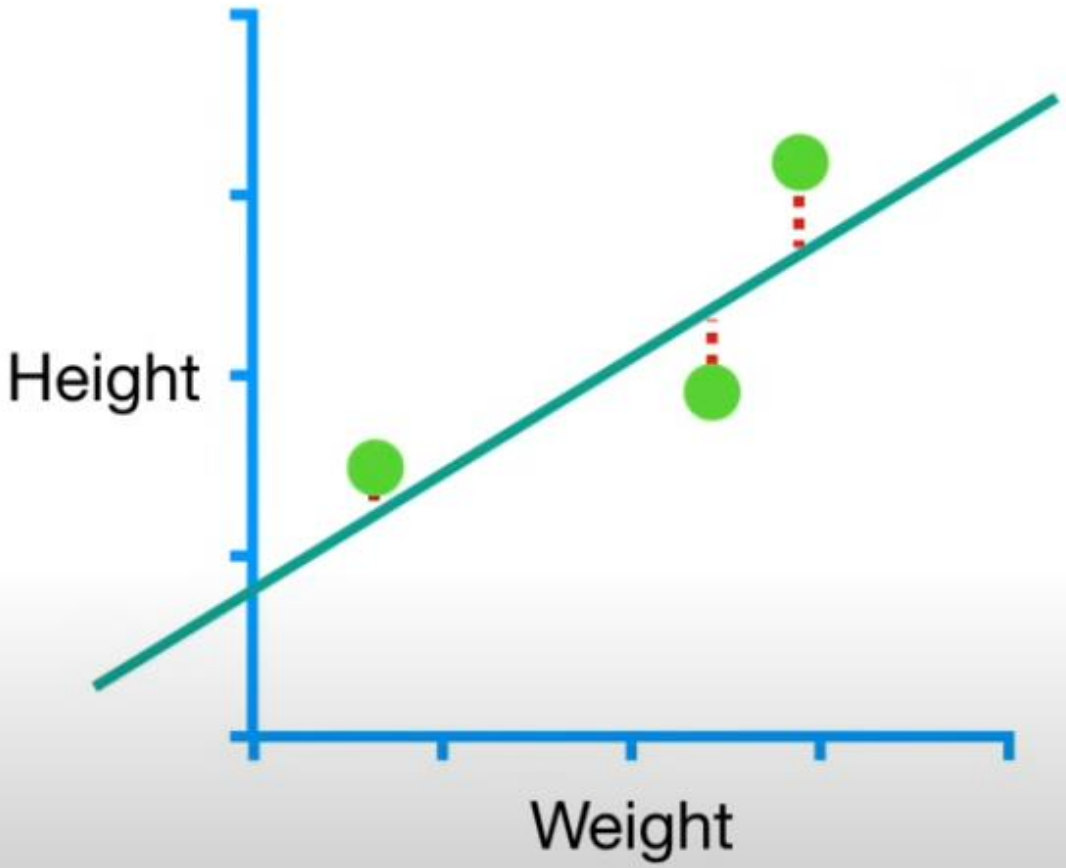
Step 4: Calculate the Step Sizes: **Step Size** = **Slope** × **Learning Rate**

Step 5: Calculate the New Parameters:

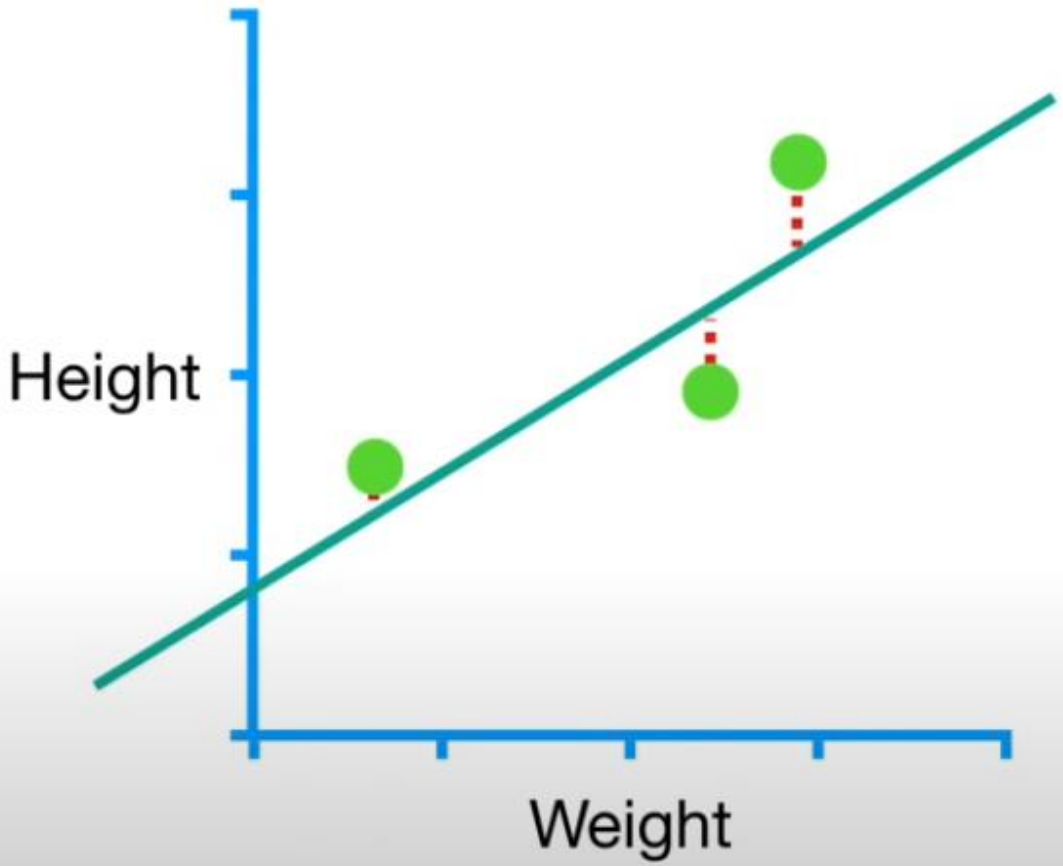
New Parameter = **Old Parameter** - **Step Size**



In our example, we only had three data points, so the math didn't take very long…



So there is a thing called
Stochastic Gradient Descent
that uses a randomly selected
subset of the data at every step
rather than the full dataset.



So there is a thing called **Stochastic Gradient Descent** that uses a randomly selected subset of the data at every step rather than the full dataset.

This reduces the time spent calculating the derivatives of the **Loss Function**.

That's all.

The End!!!