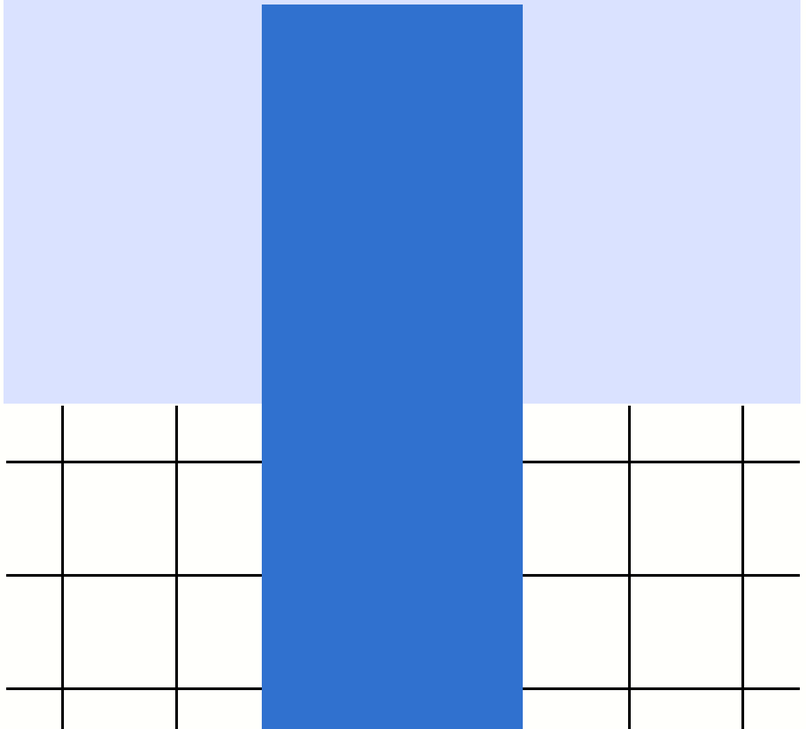
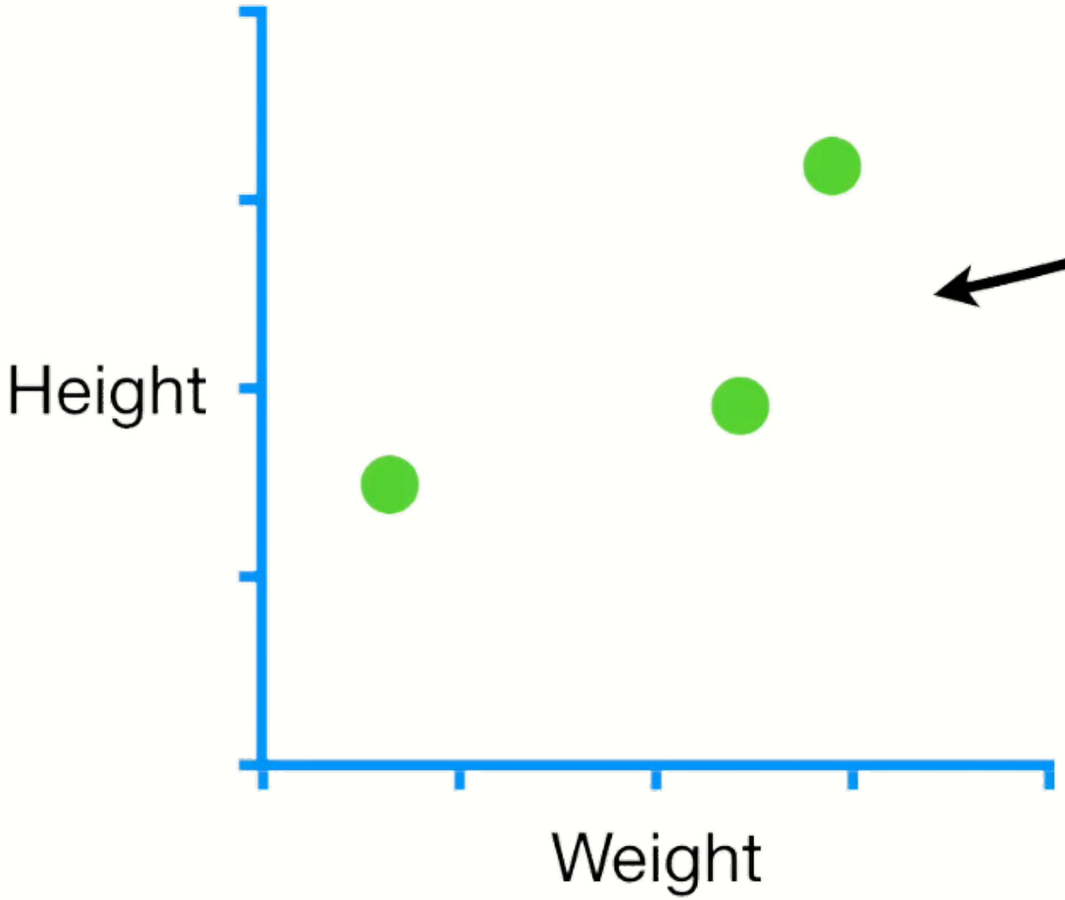
GRADIENT DESCENT & STOCHASTIC GRADIENT DESCENT

– Matee Vadrukchid –

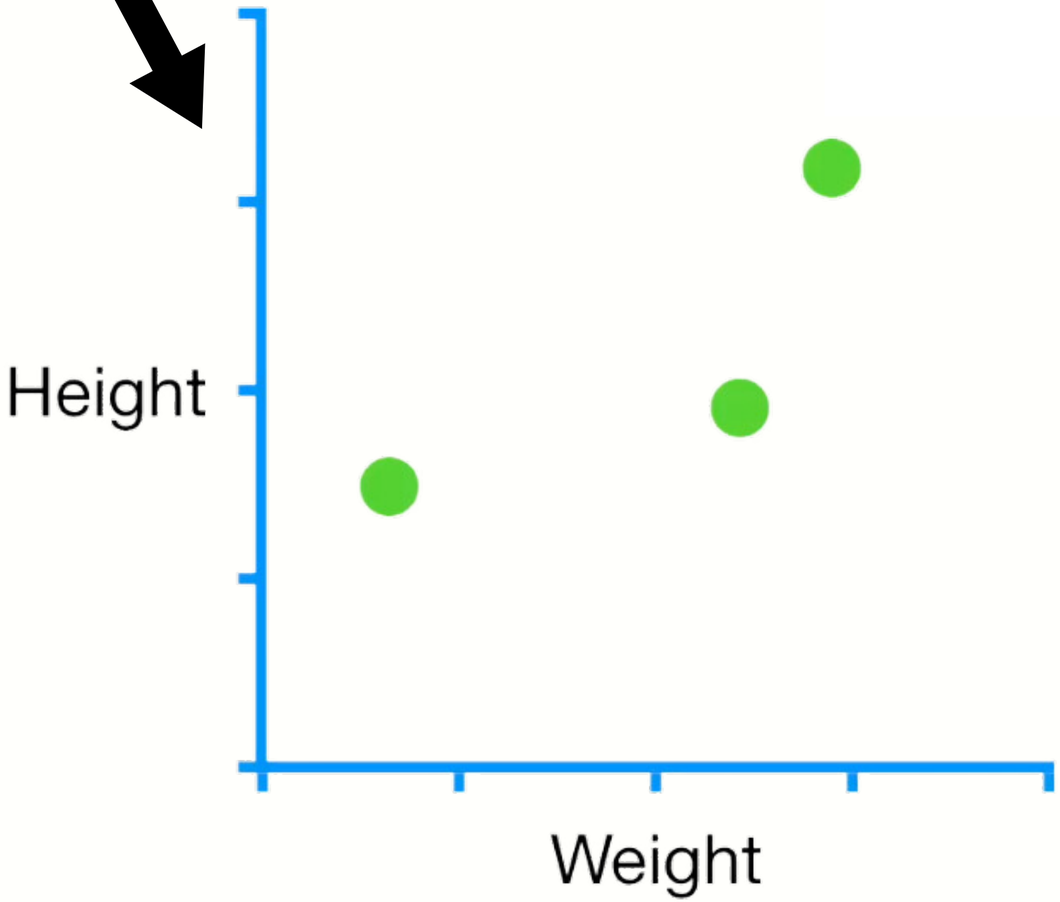


In Statistics, Machine Learning and other Data Science fields,
we optimize a lot of stuff.

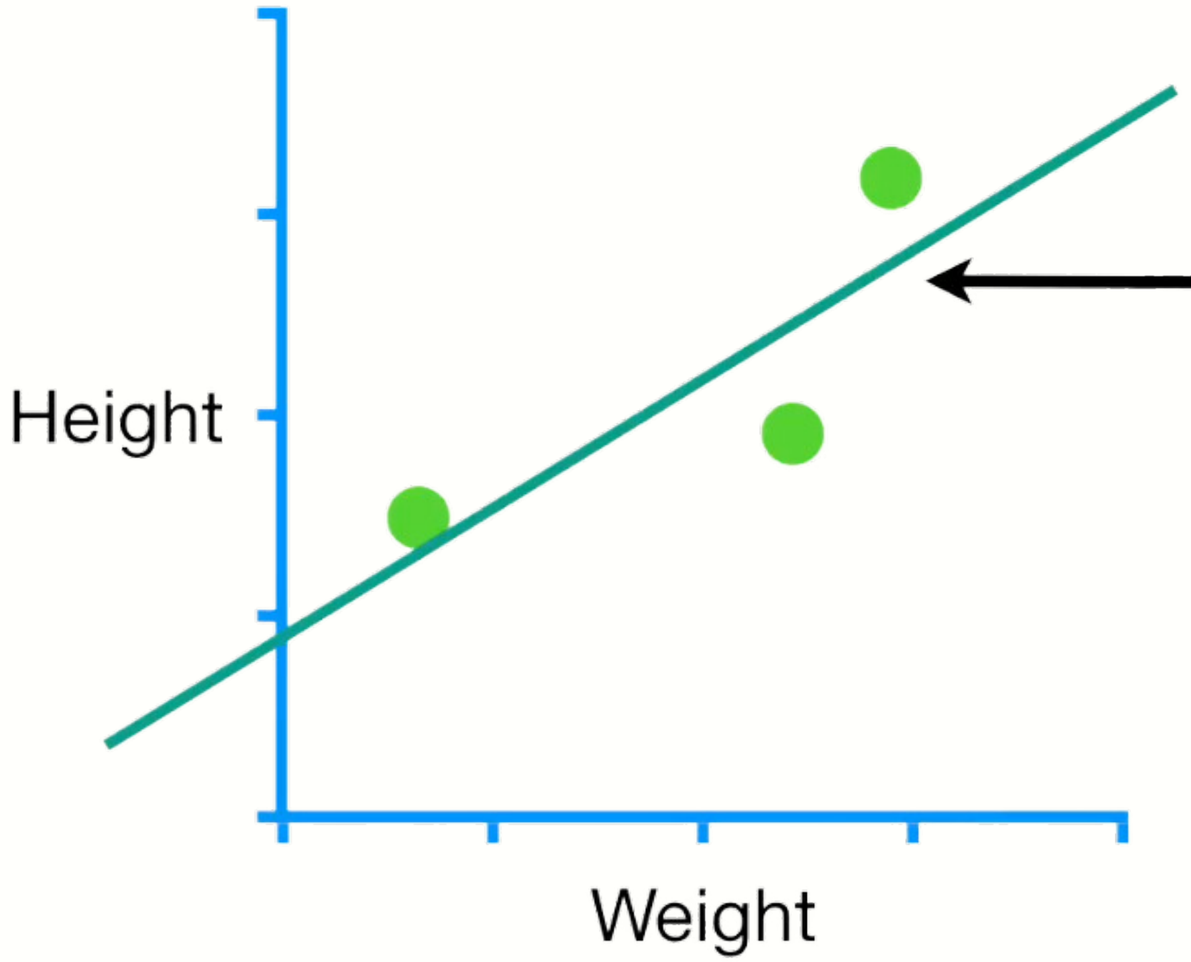
So let’s start with a simple data set.



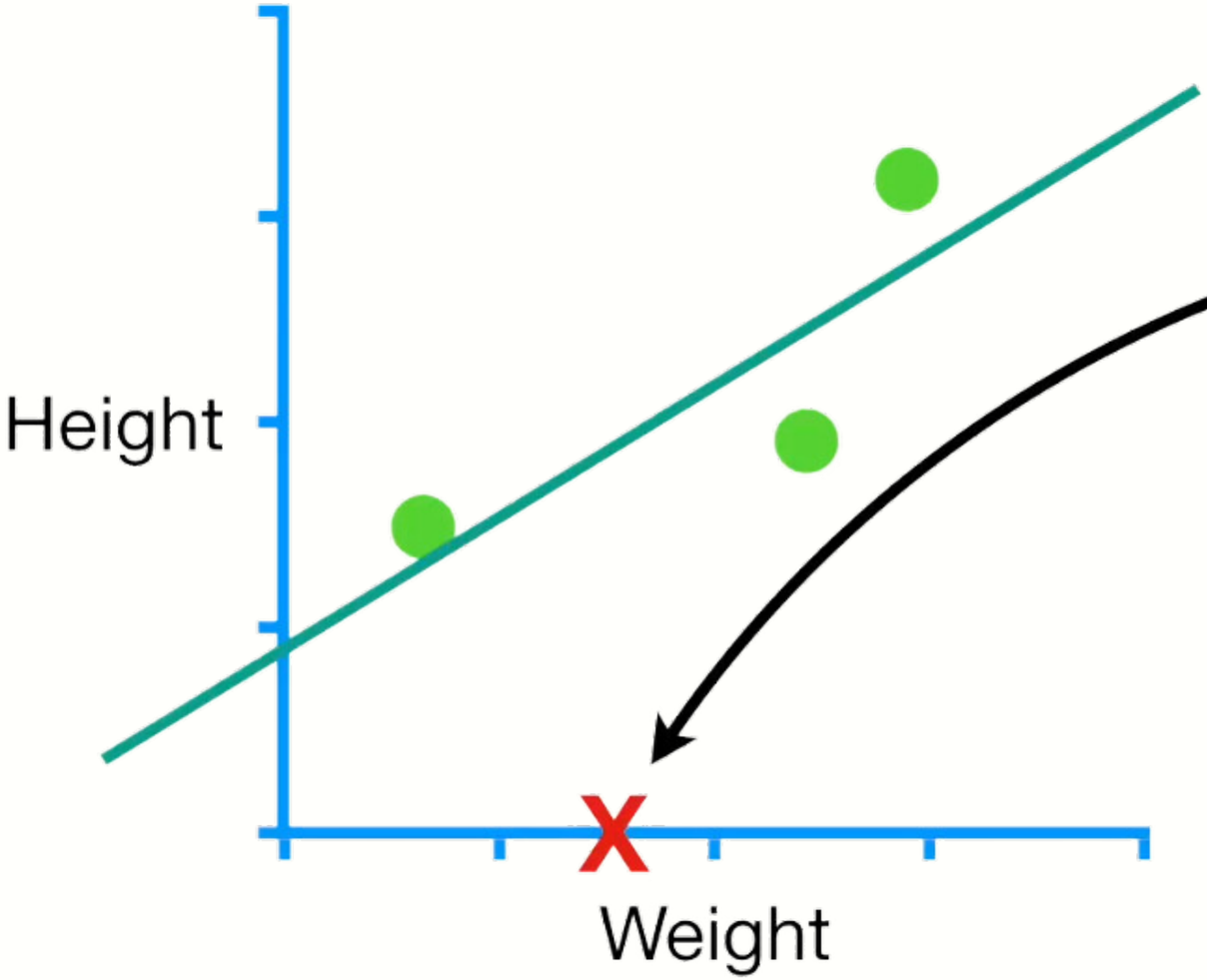
...and on the y-axis we have Height.



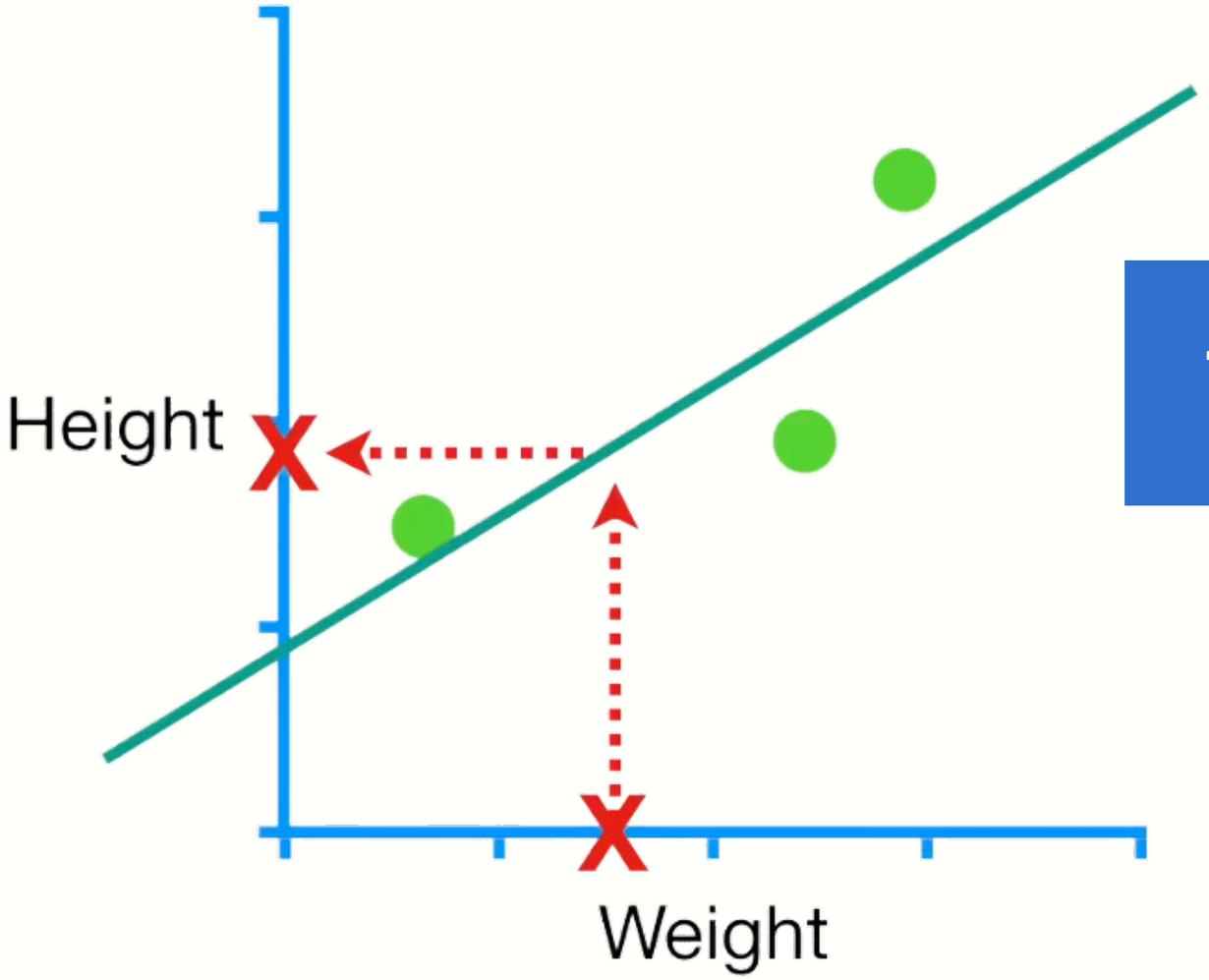
On the x-axis, we have Weight…



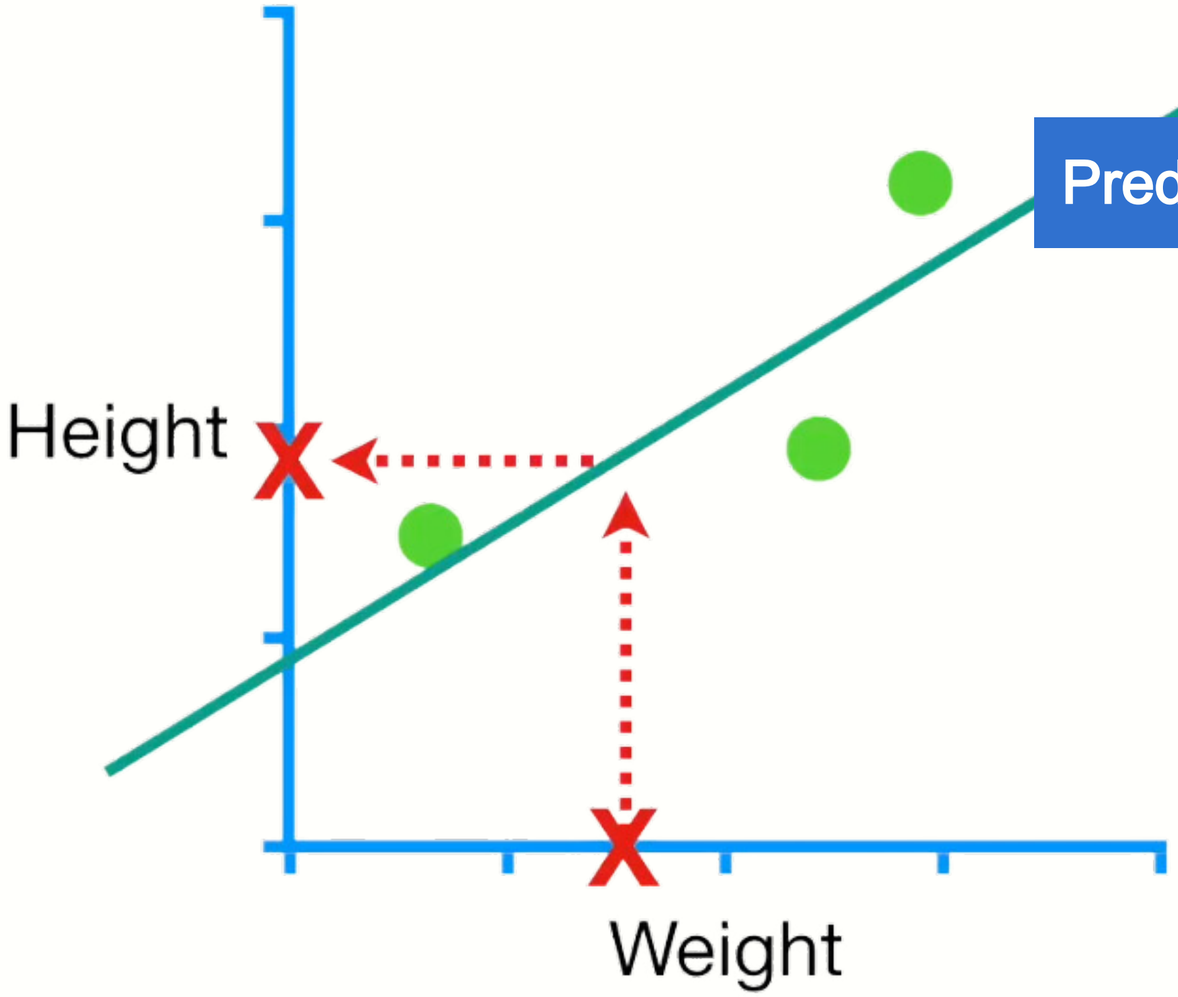
If we fit a line to the data…



...and someone tells us that they weigh 1.5...

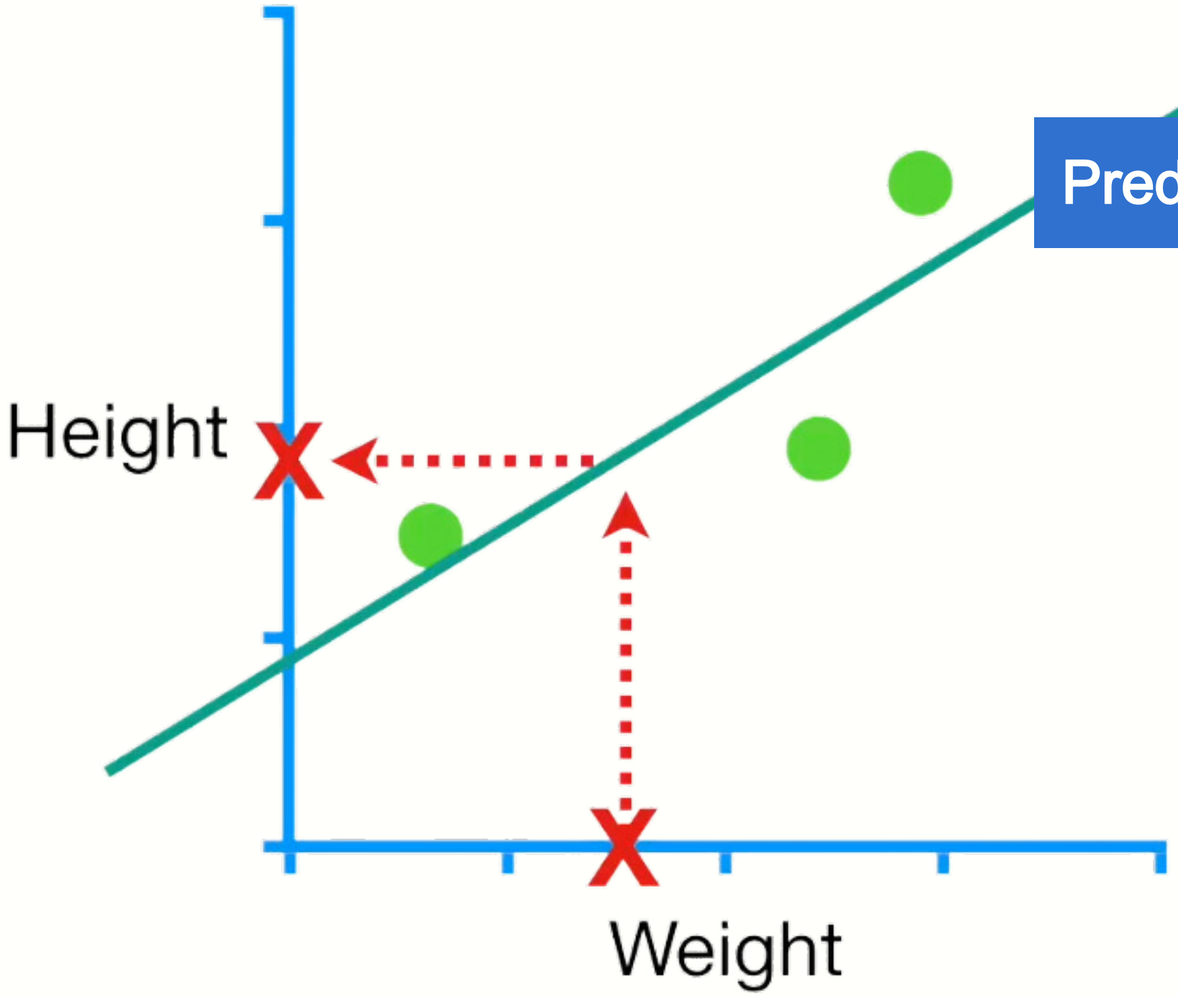


...we can use the line to predict that they will be 1.9 tall.



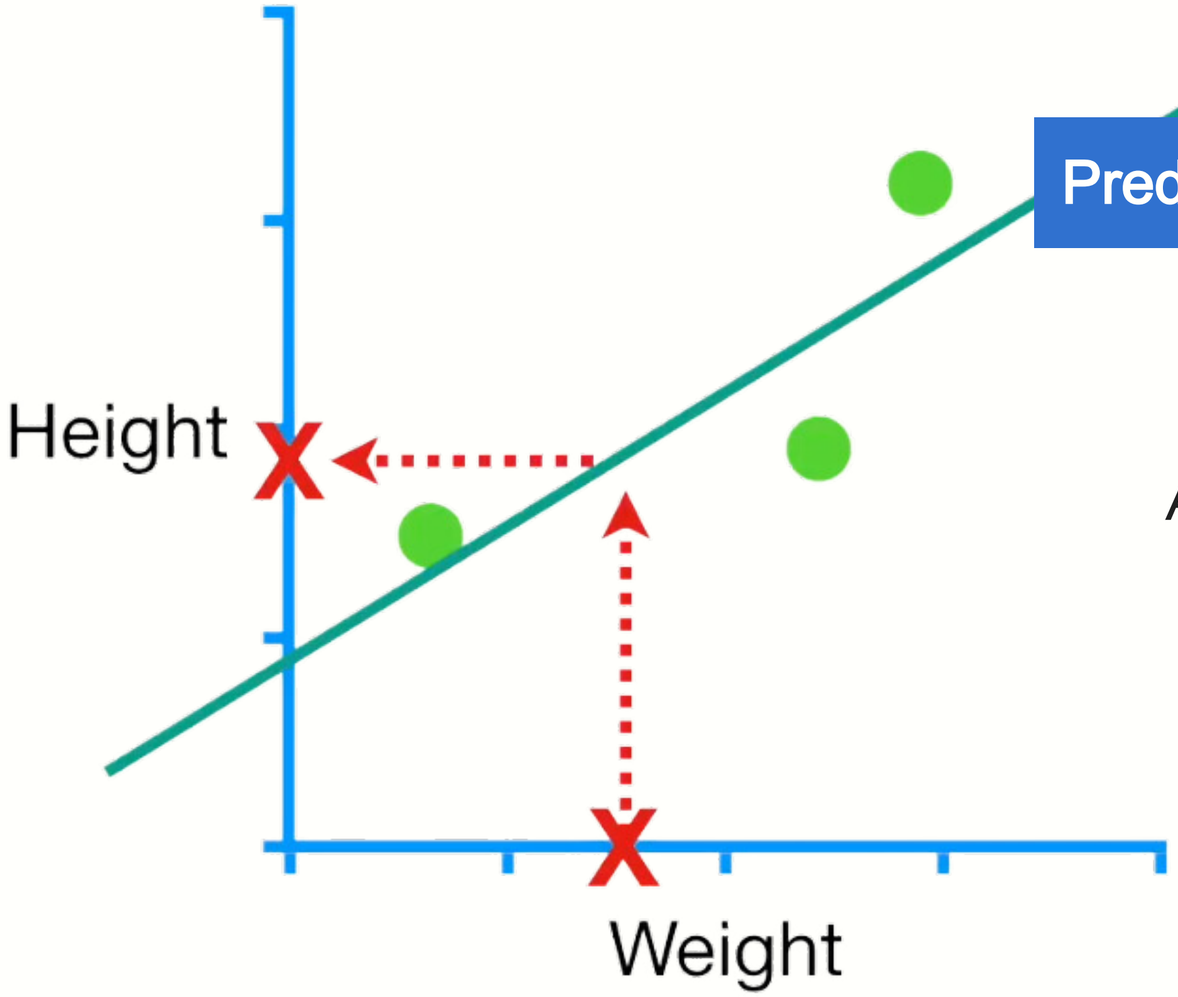
Predicted Height = intercept + slope × Weight

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



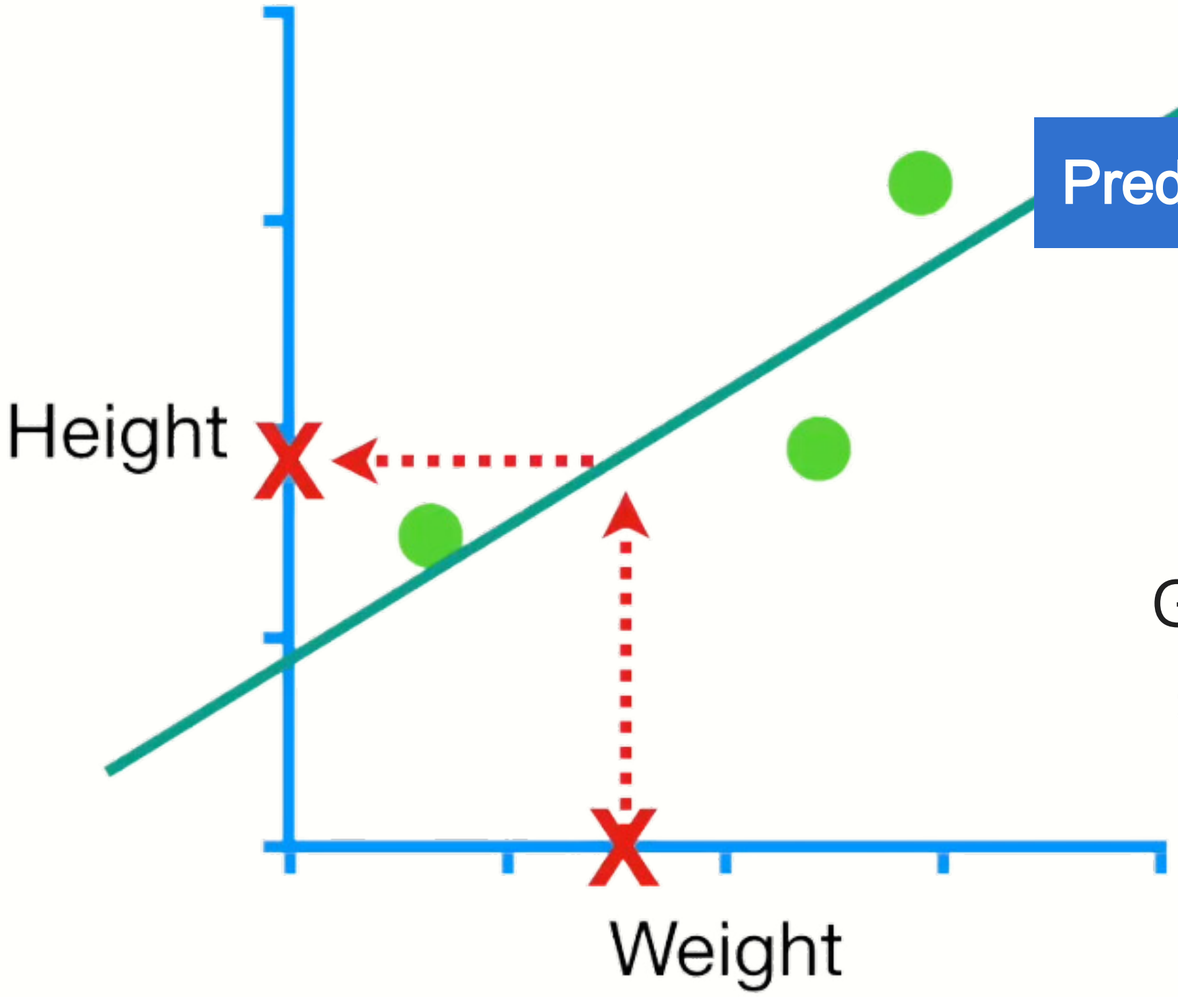
Predicted Height = intercept + slope × Weight

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



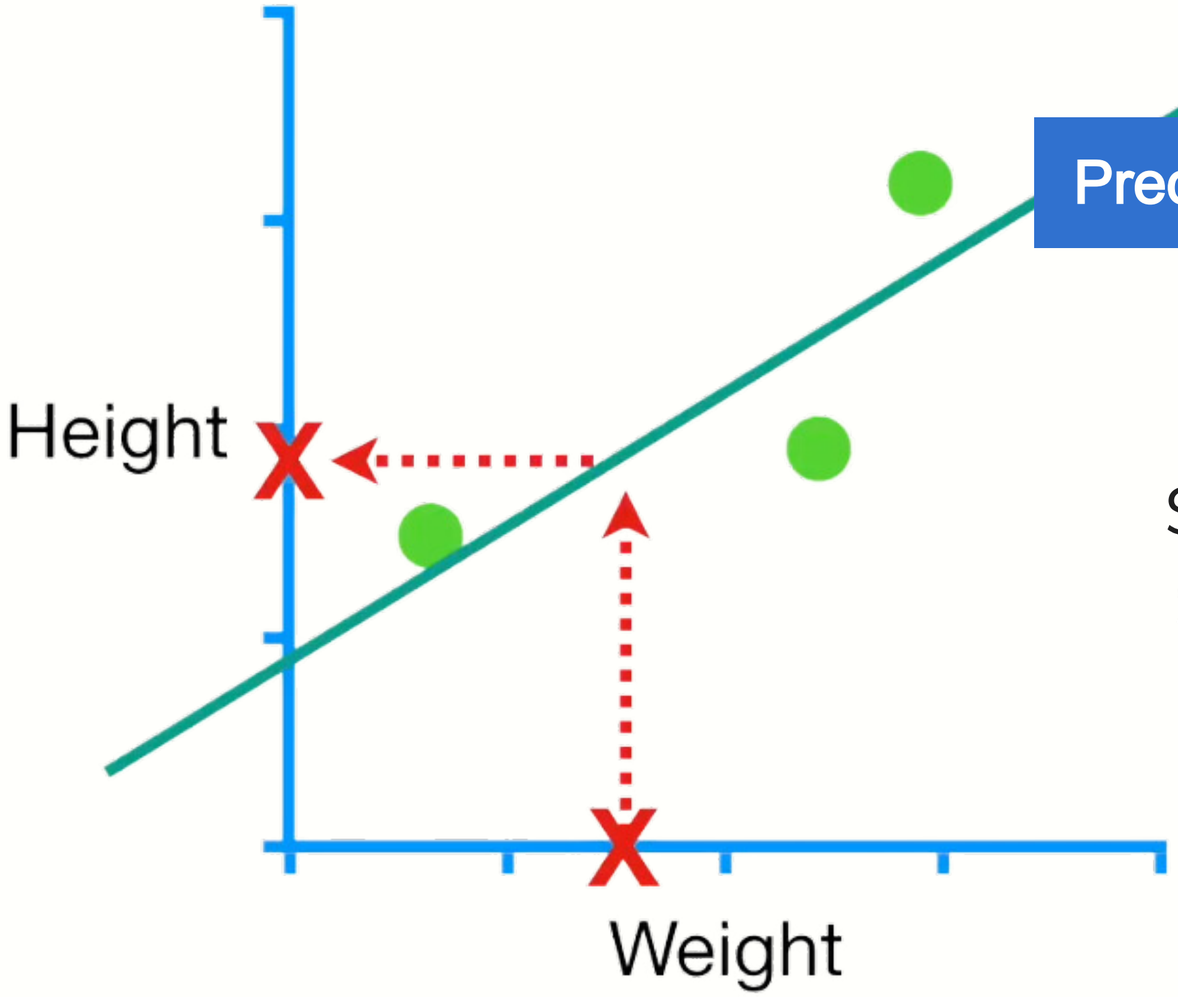
Predicted Height = intercept + slope × Weight

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



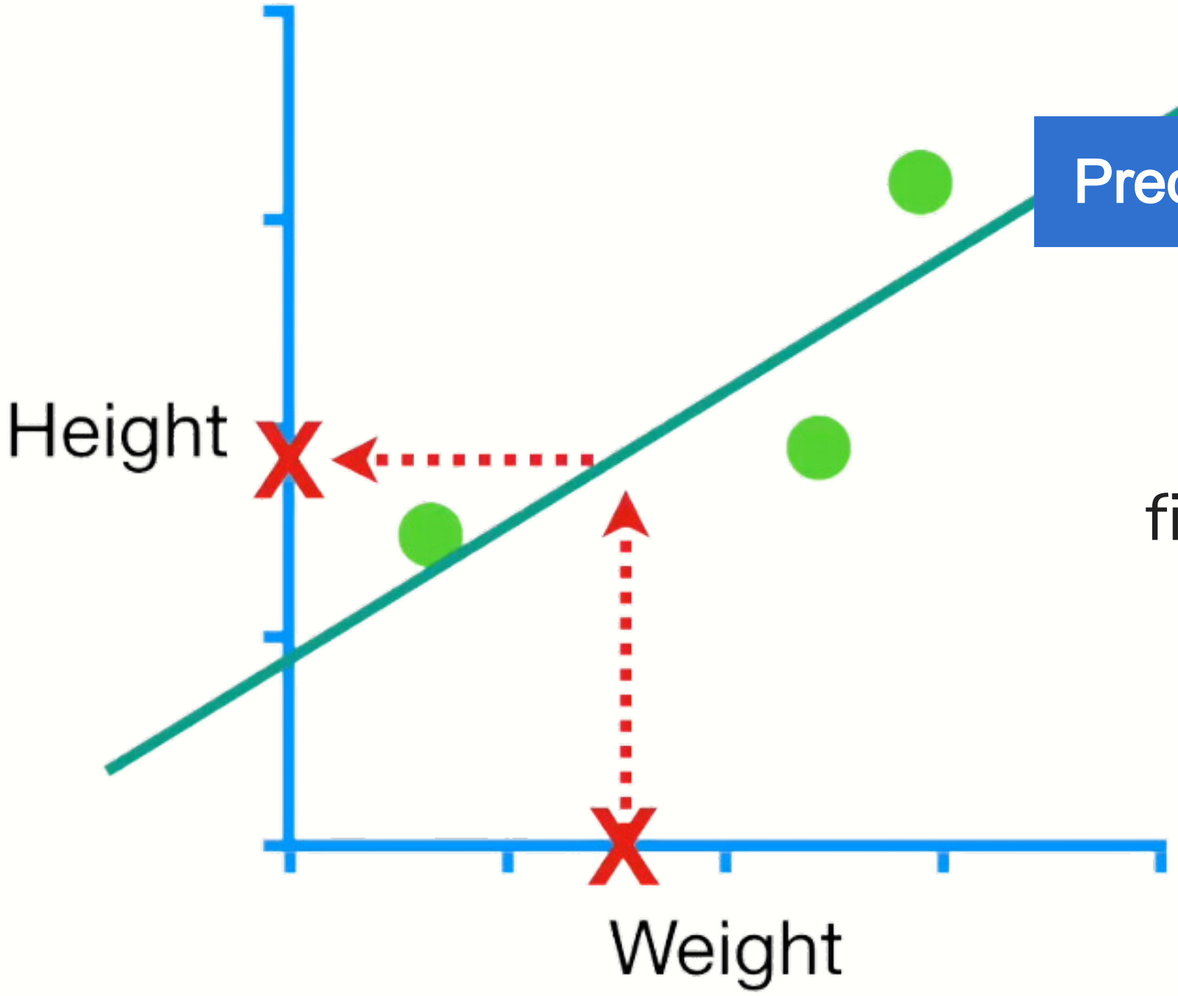
Predicted Height = intercept + slope × Weight

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



Predicted Height = intercept + 0.64 × Weight

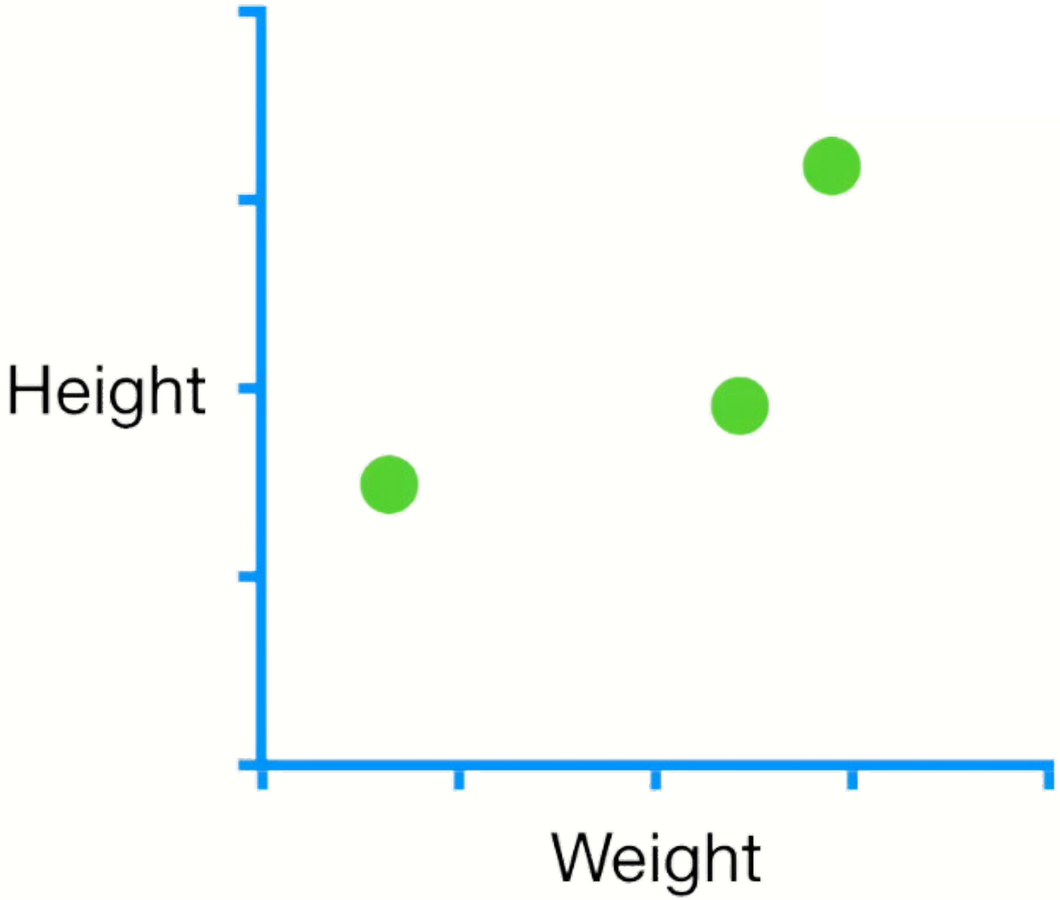
So for now, let's just plug in the Least Squares estimate for the Slope, 0.64.



Predicted Height = intercept + 0.64 × Weight

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

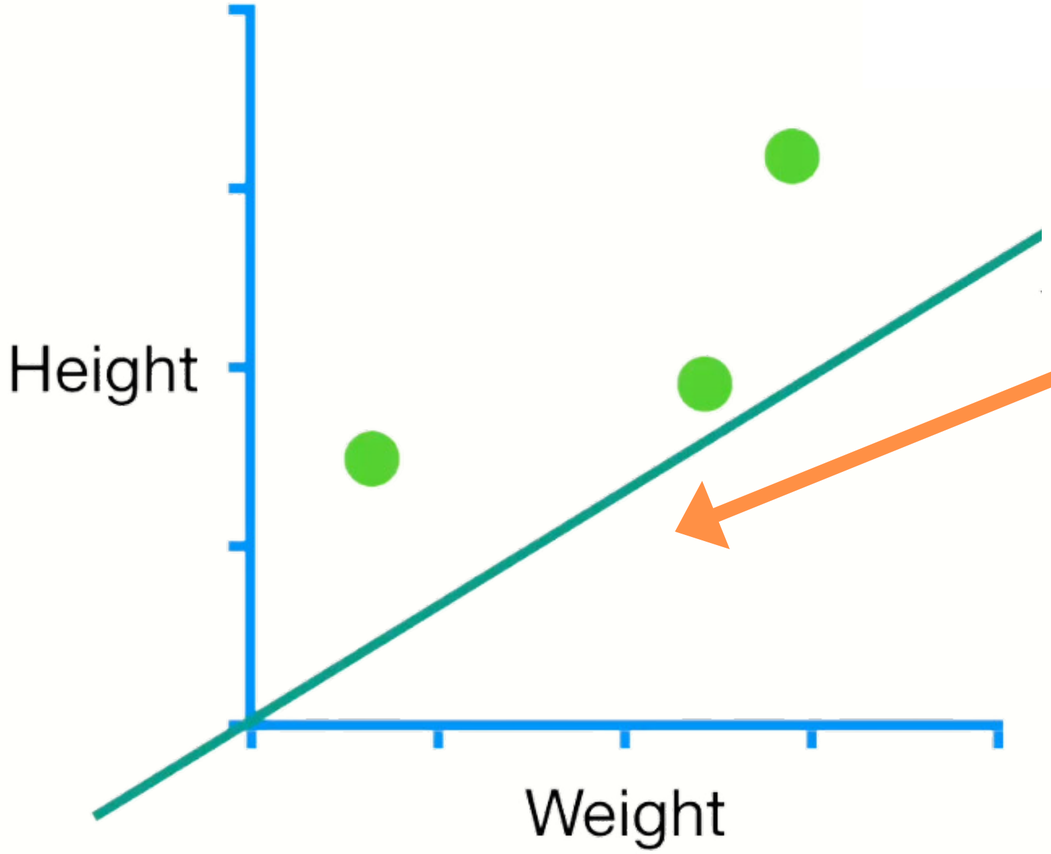
Predicted Height = intercept + 0.64 × Weight



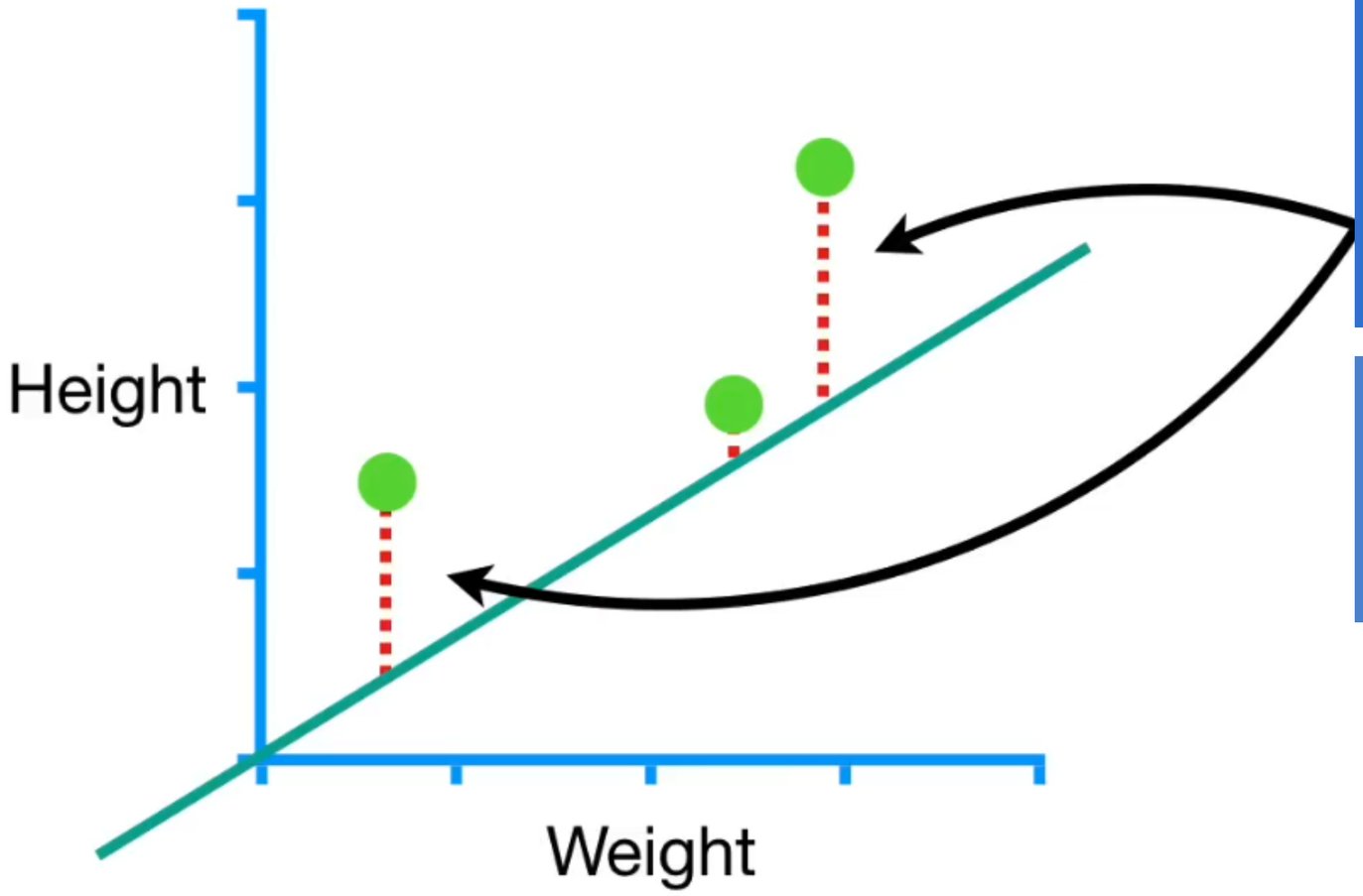
The first thing we do is pick a random value for the Intercept.

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

Predicted Height = 0 + 0.64 × Weight

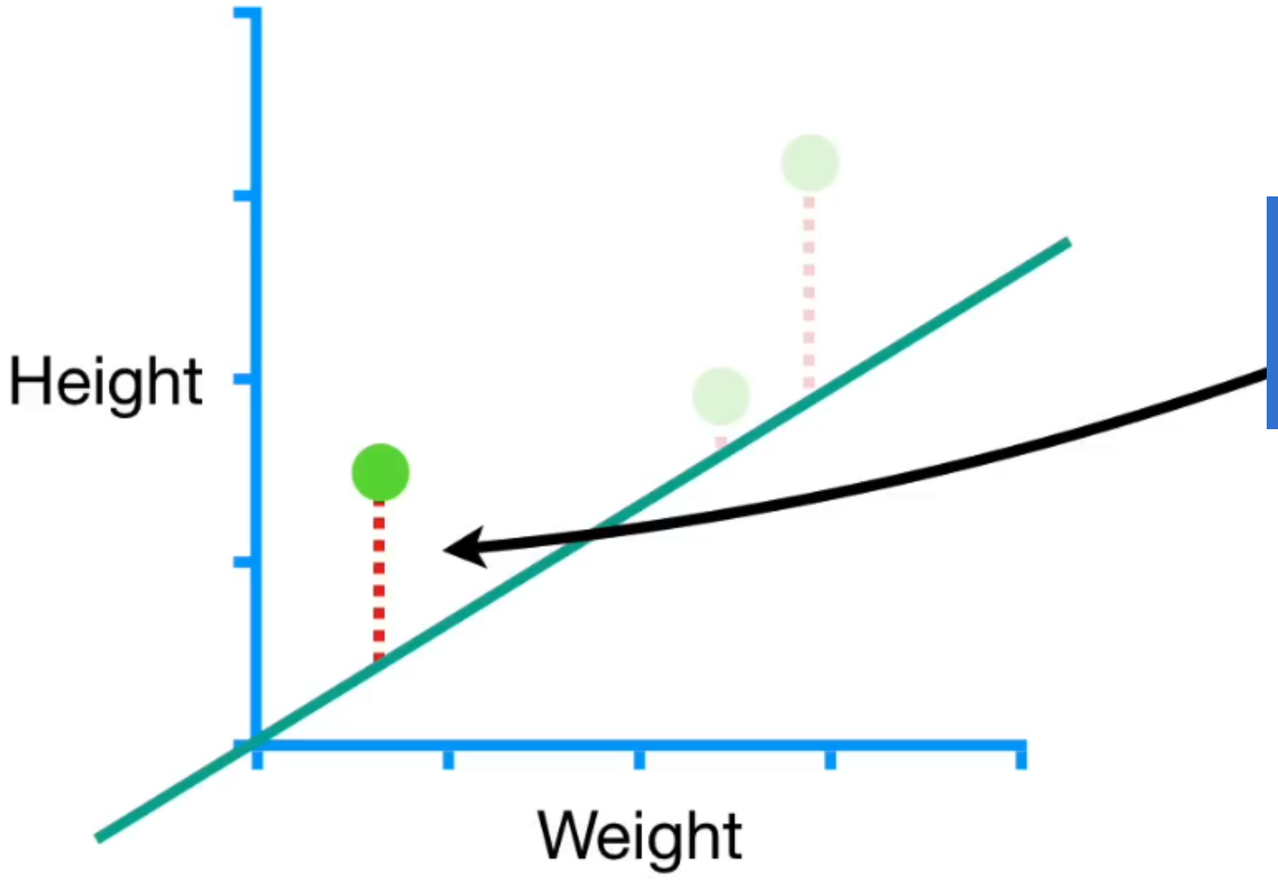


In this case, we'll use 0, but any number will do and that gives us the equation for this line.

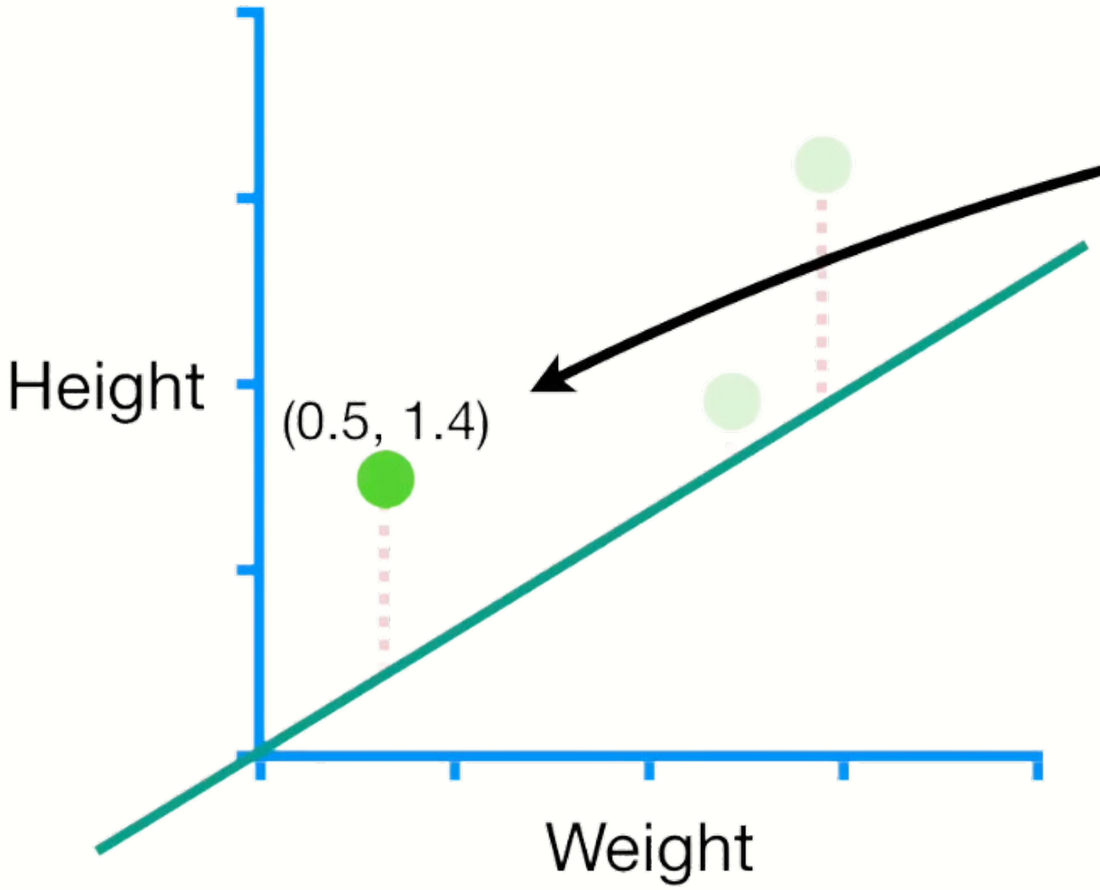


In this example, we will evaluate how well this line fits the data with the Sum of the Squared Residuals.

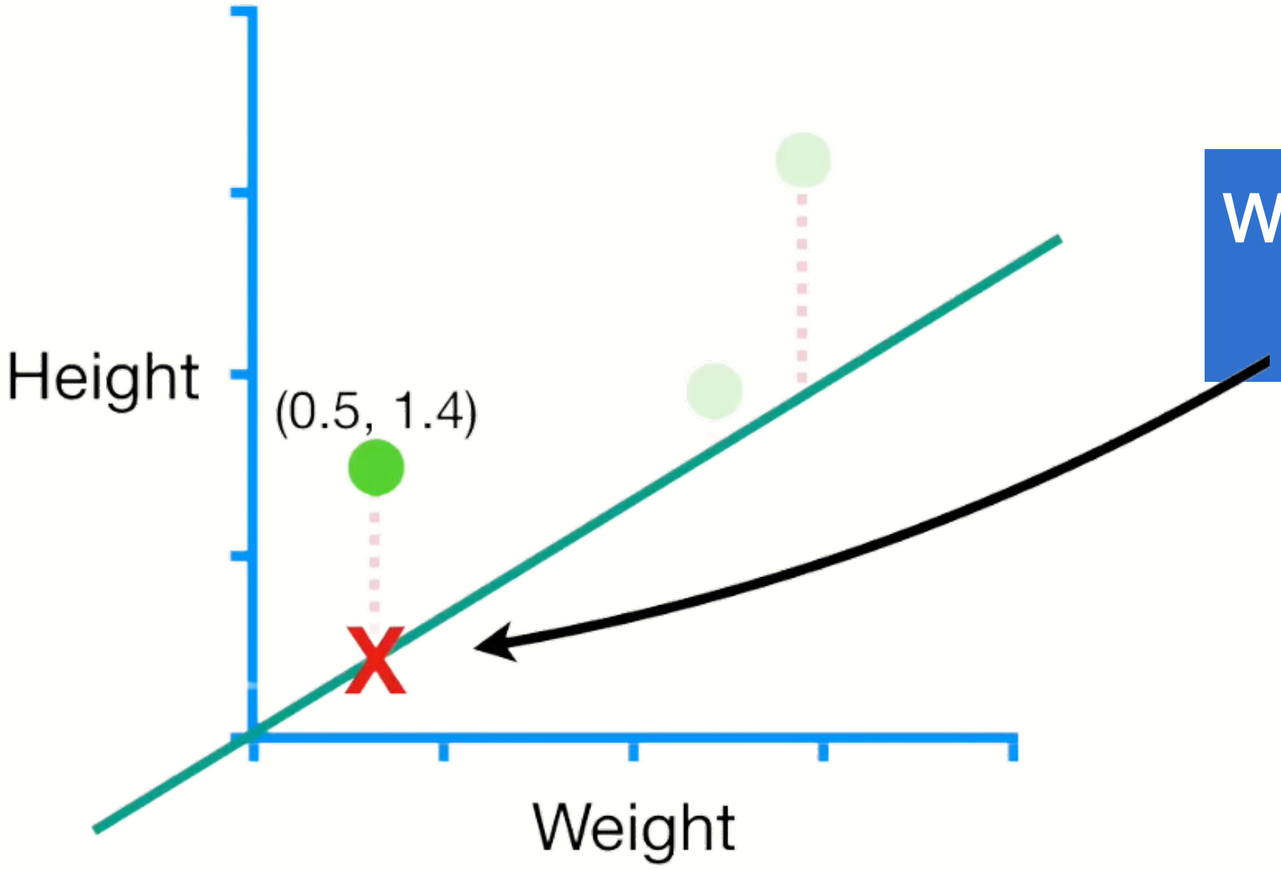
NOTE: In Machine Learning lingo, The Sum of the Squared Residuals is a type of Loss Function.



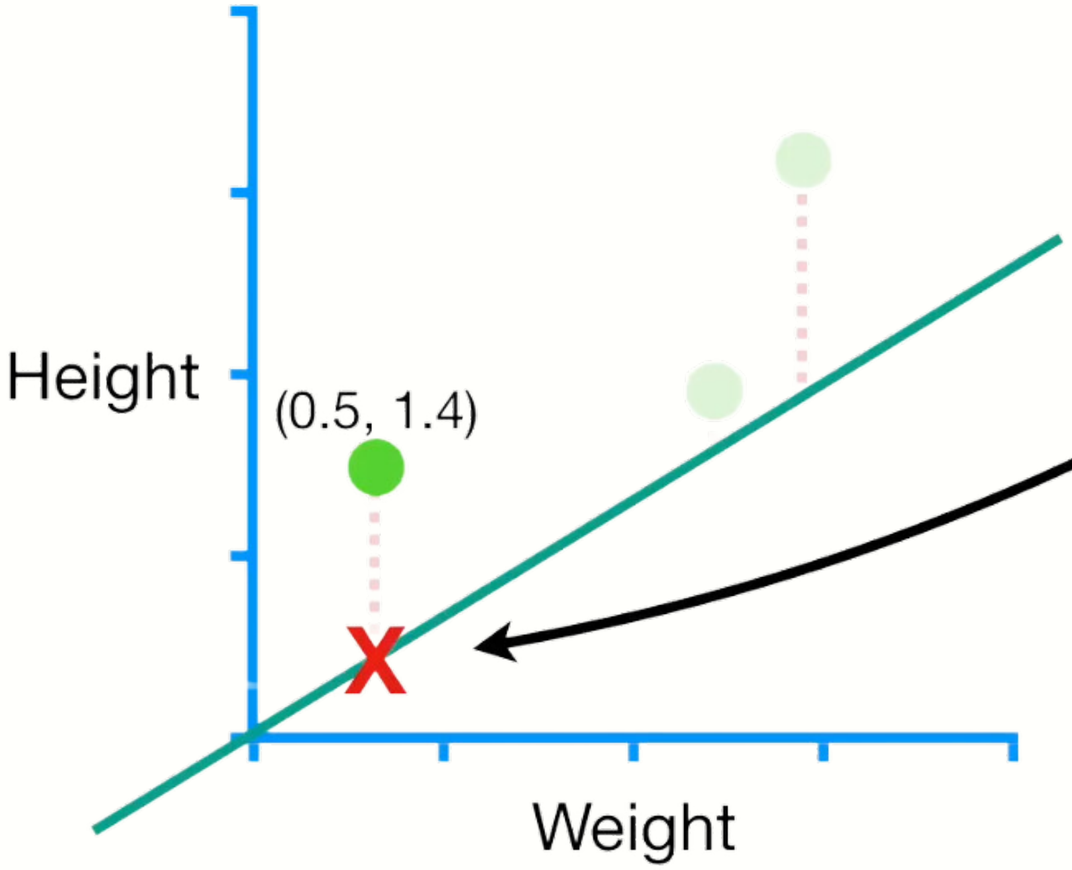
We'll start by calculating this residual.



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



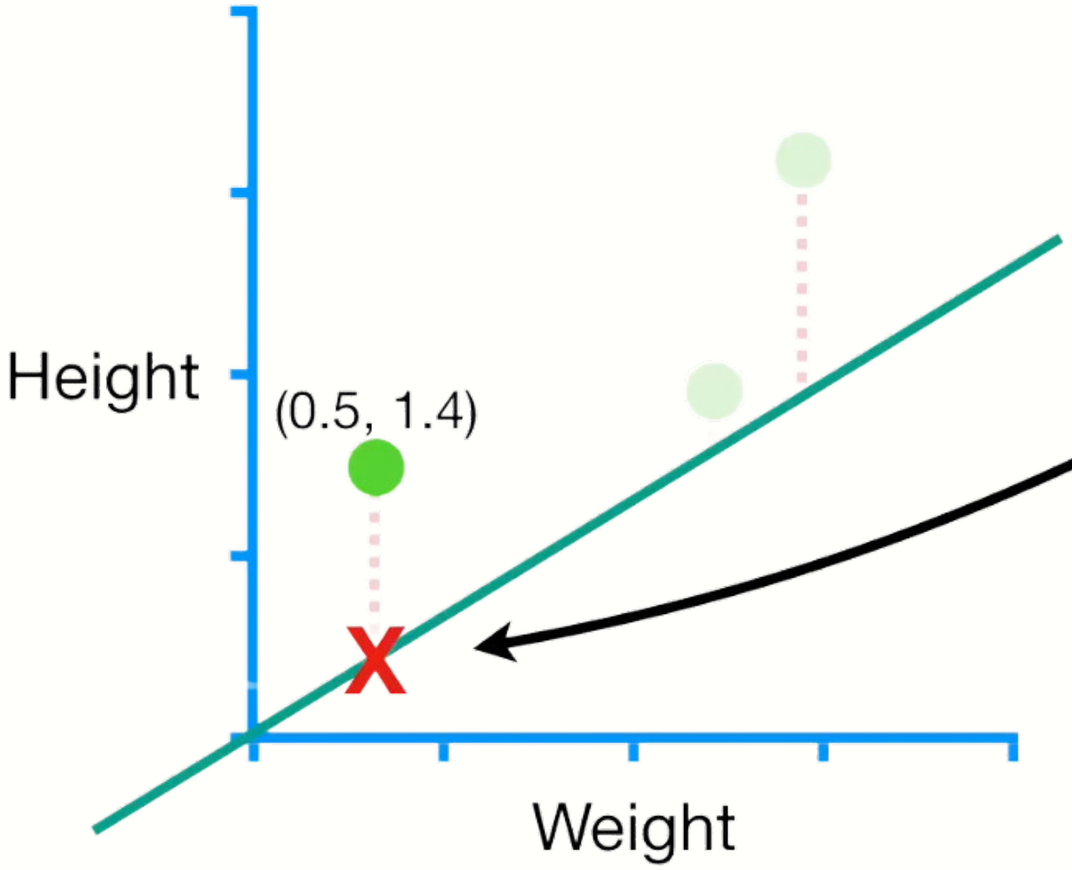
We get the Predicted Height, the point on the line…



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

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

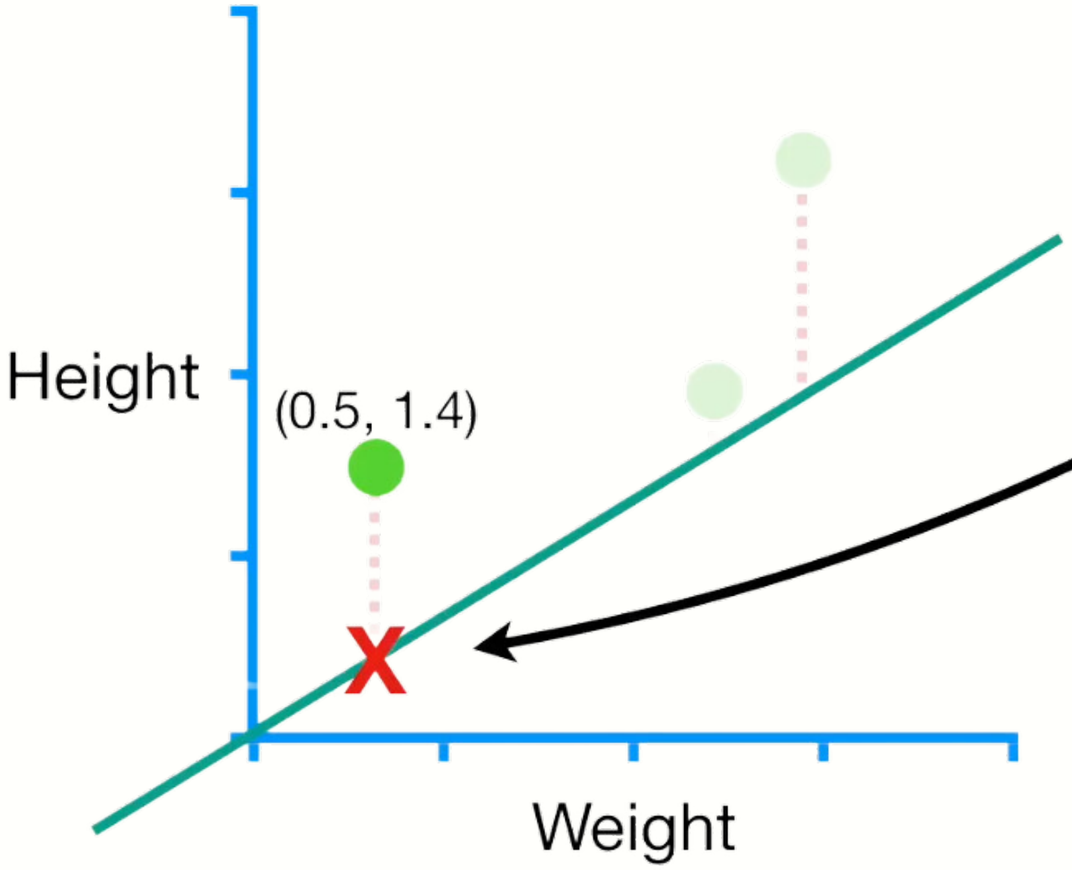
Predicted Height = 0 + 0.64 × Weight



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

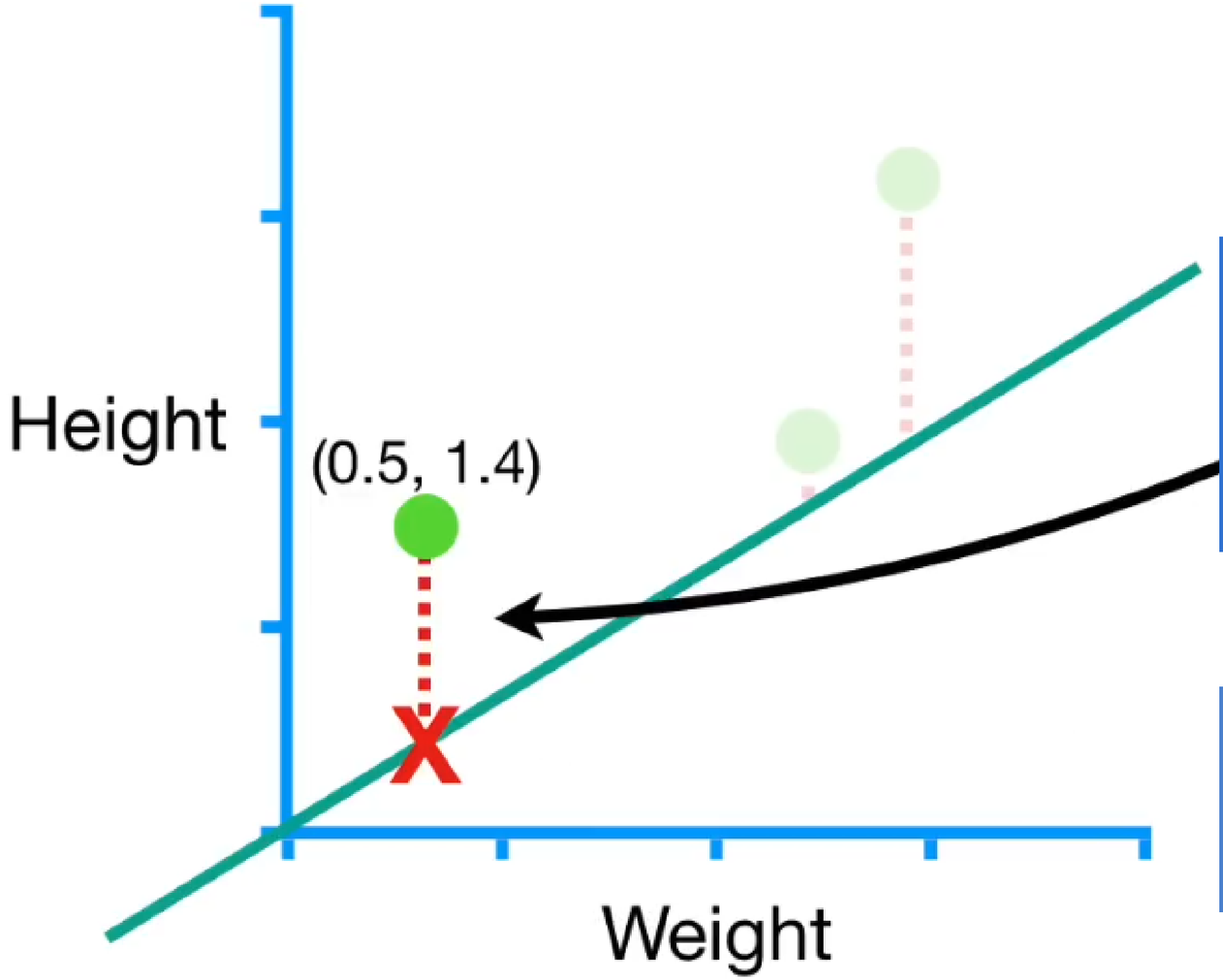
…by plugging Weight = 0.5 into the equation for the line…

Predicted Height = 0 + 0.64 × 0.5



...and the Predicted Height is 0.32.

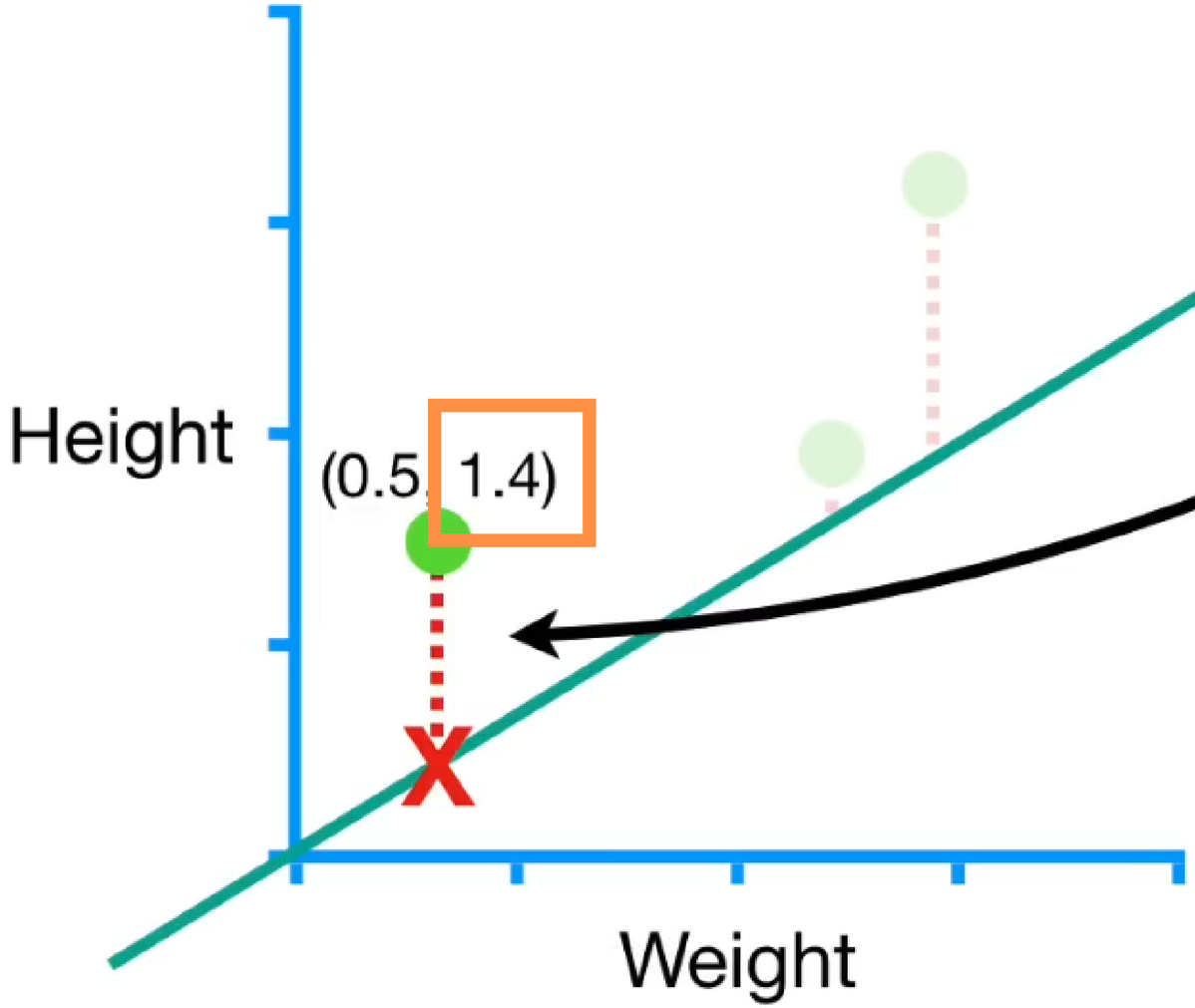
Predicted Height = 0 + 0.64 × 0.5
= 0.32



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

Predicted Height = 0 + 0.64 × 0.5
= 0.32

Residual = Observed Height - Predicted Height



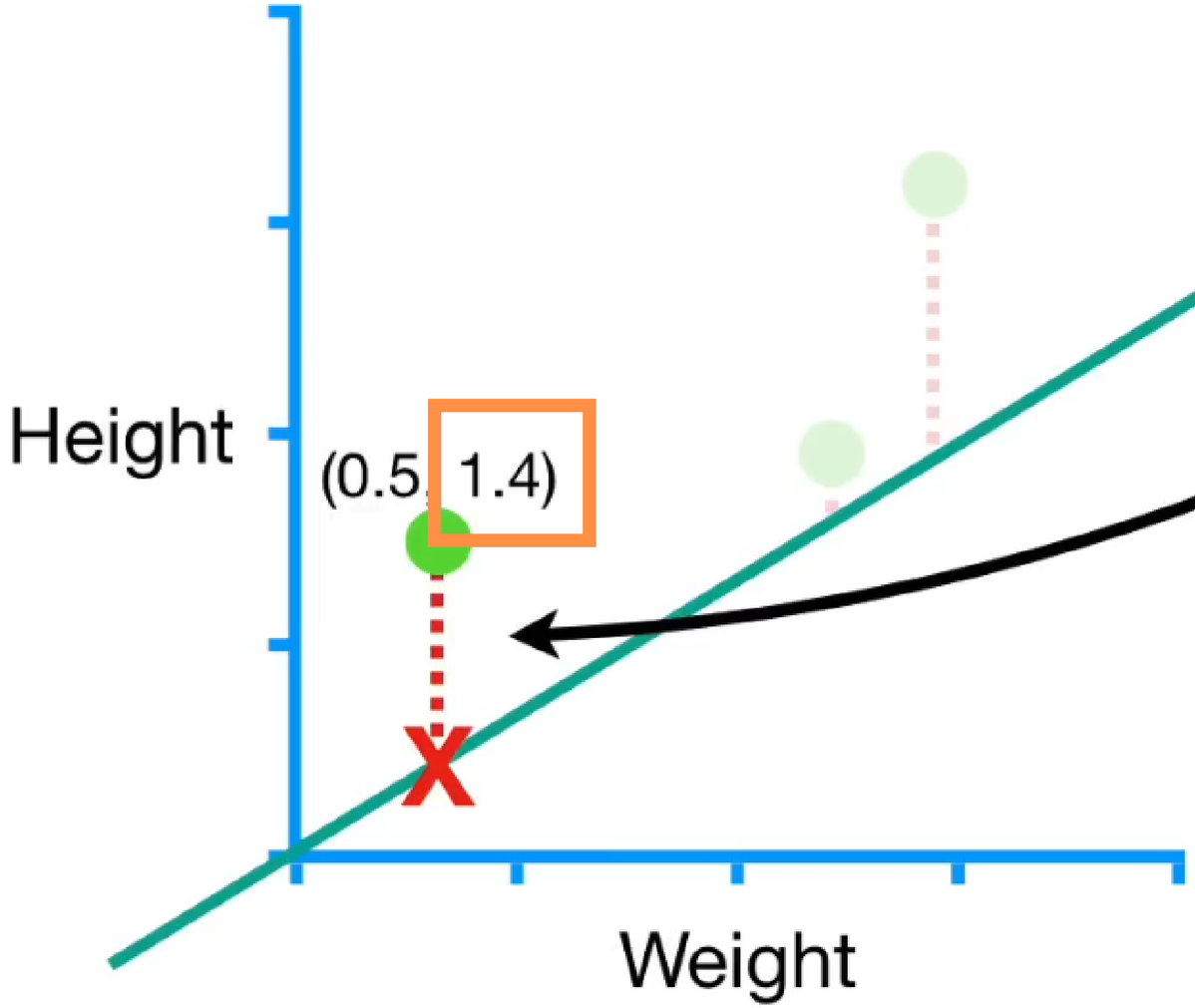
The residual is the difference between the Observed Height and the Predicted Height..

Predicted Height = 0 + 0.64 × 0.5
= 0.32

Residual = Observed Height - Predicted Height

= 1.4 - 0.32

= 1.1



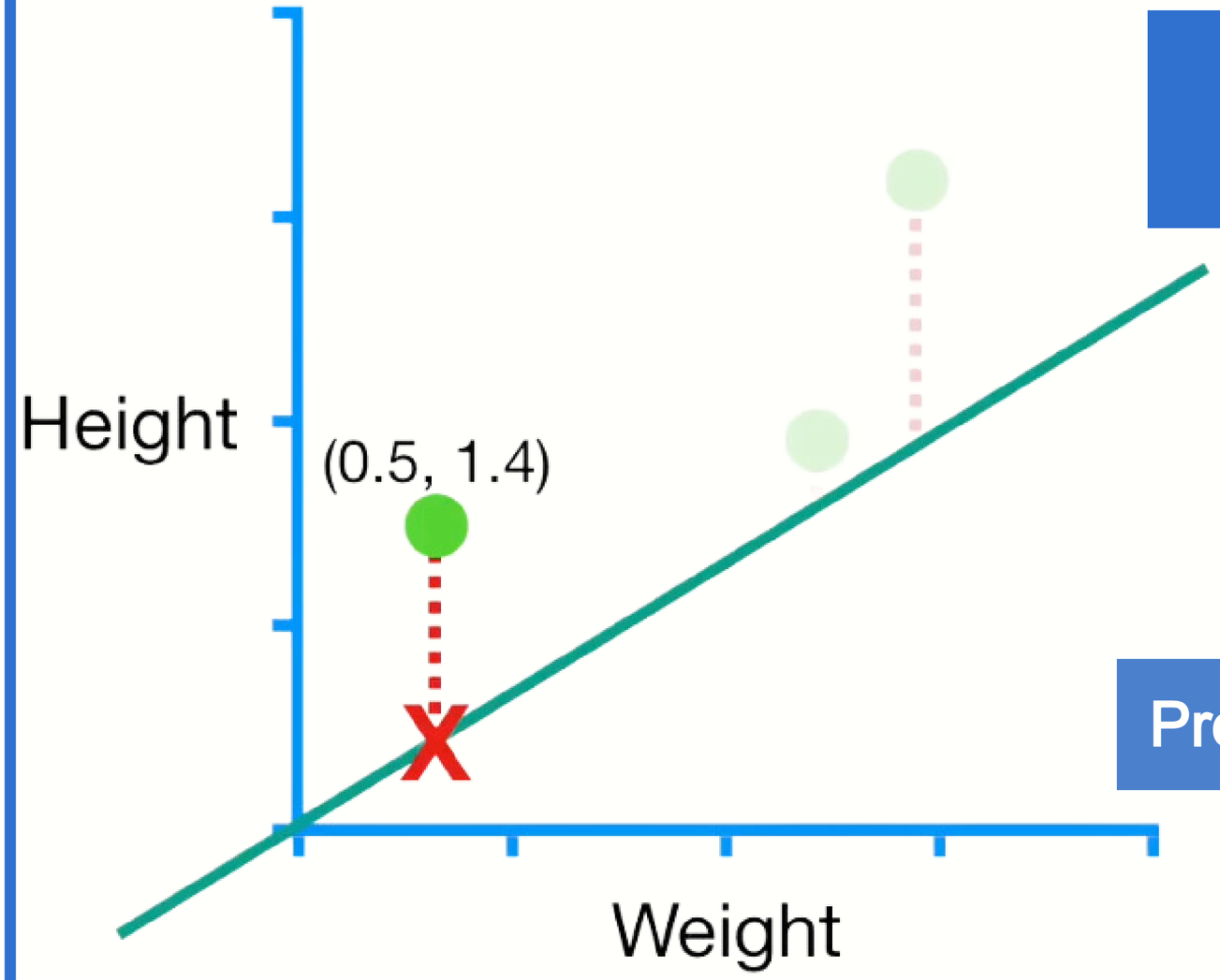
..and that gives us 1.1 for the residual.

Predicted Height = 0 + 0.64 × 0.5

= 0.32

Sum of squared residuals = 1.1^2

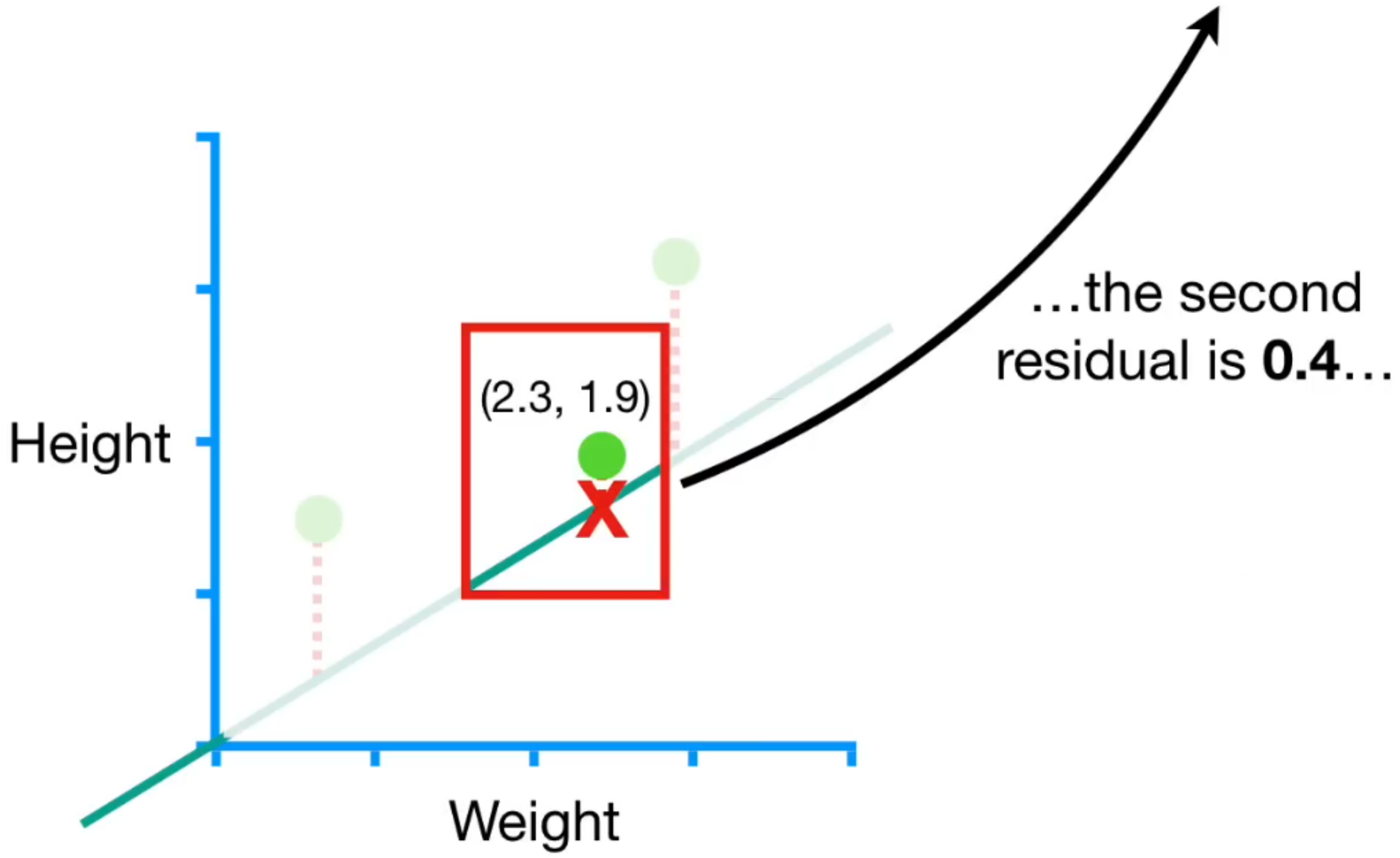
We'll keep track of the Sum of the Squared Residuals up here.



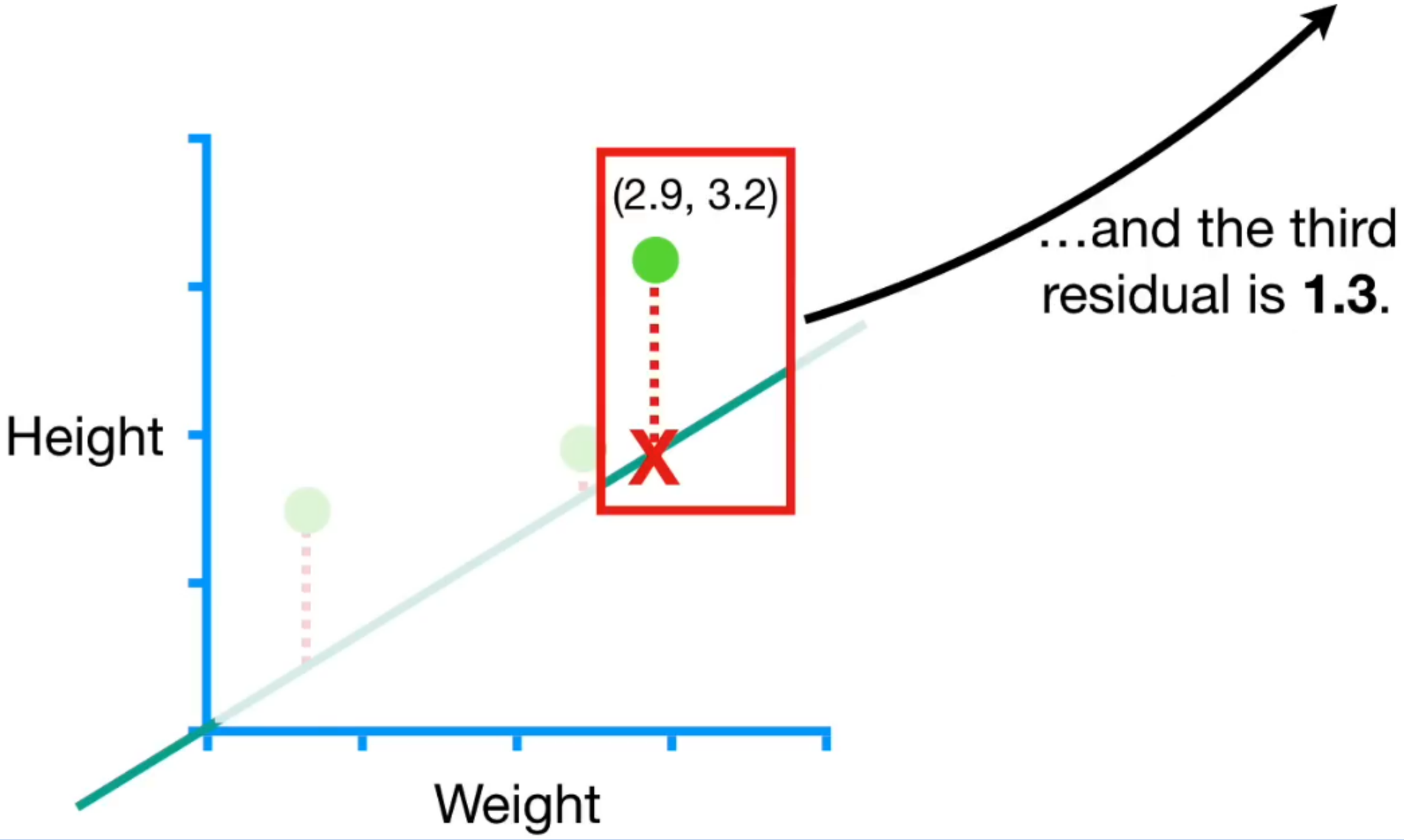
Residual = 1.1

Predicted Height = 0 + 0.64 × 0.5 = 0.32

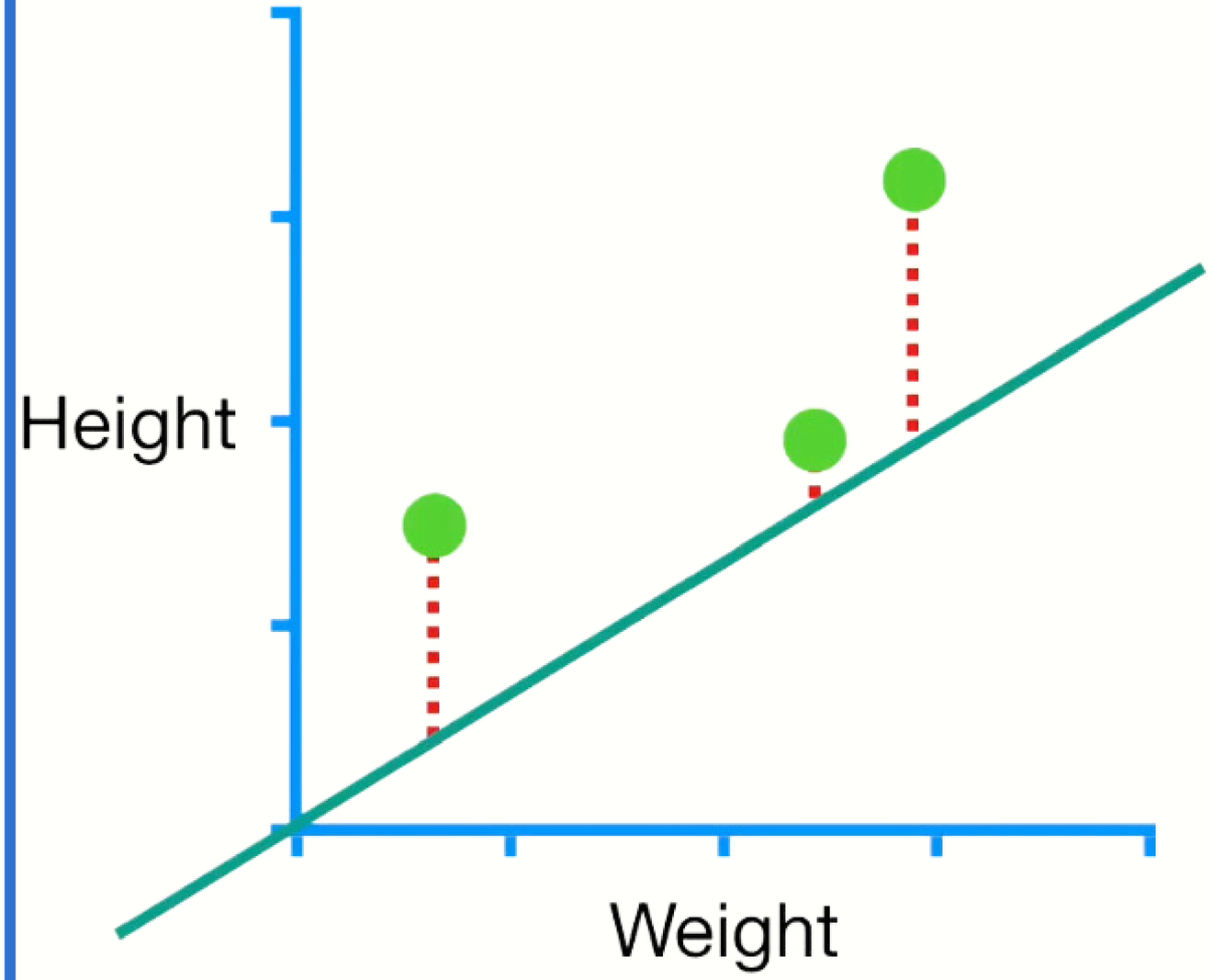
Sum of squared residuals = $1.1^2 + 0.4^2$



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

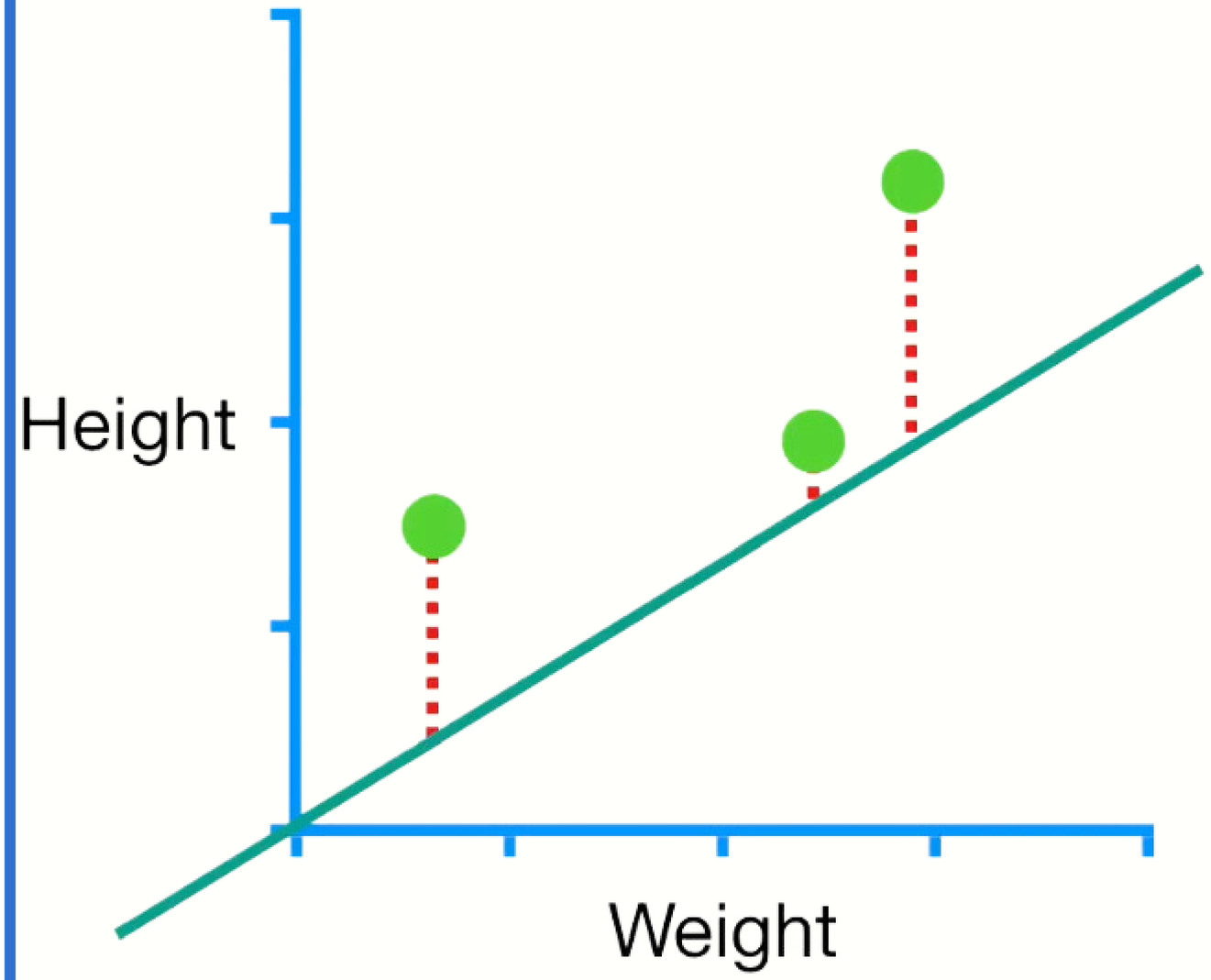


Sum of squared residuals = $1.1^2 + 0.4^2 + 1.3^2 = 3.1$

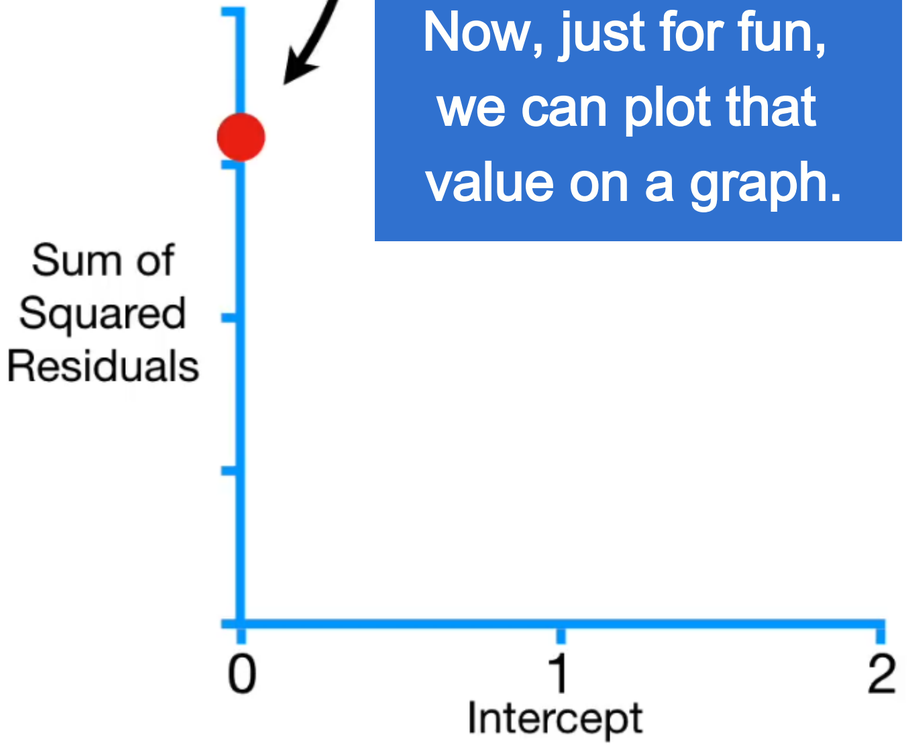


In the end, 3.1 is the Sum of the Squared Residuals.

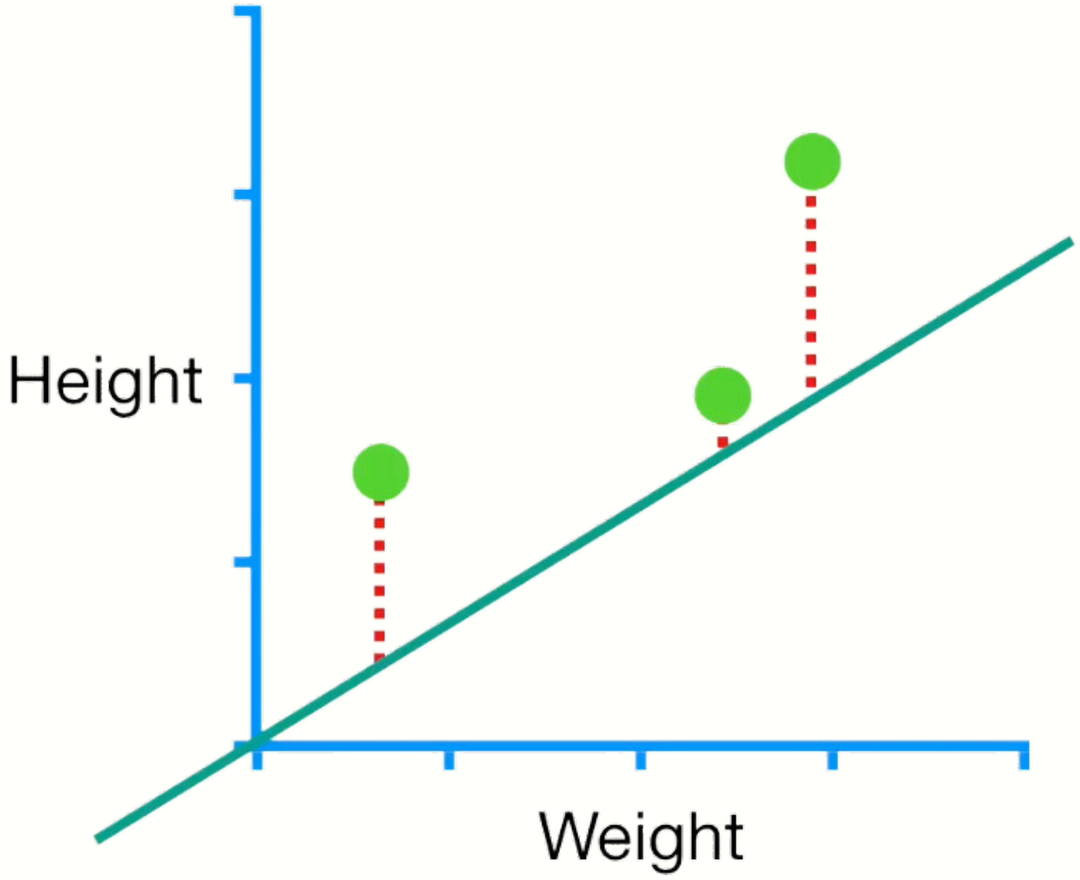
Sum of squared residuals = $1.1^2 + 0.4^2 + 1.3^2 = 3.1$



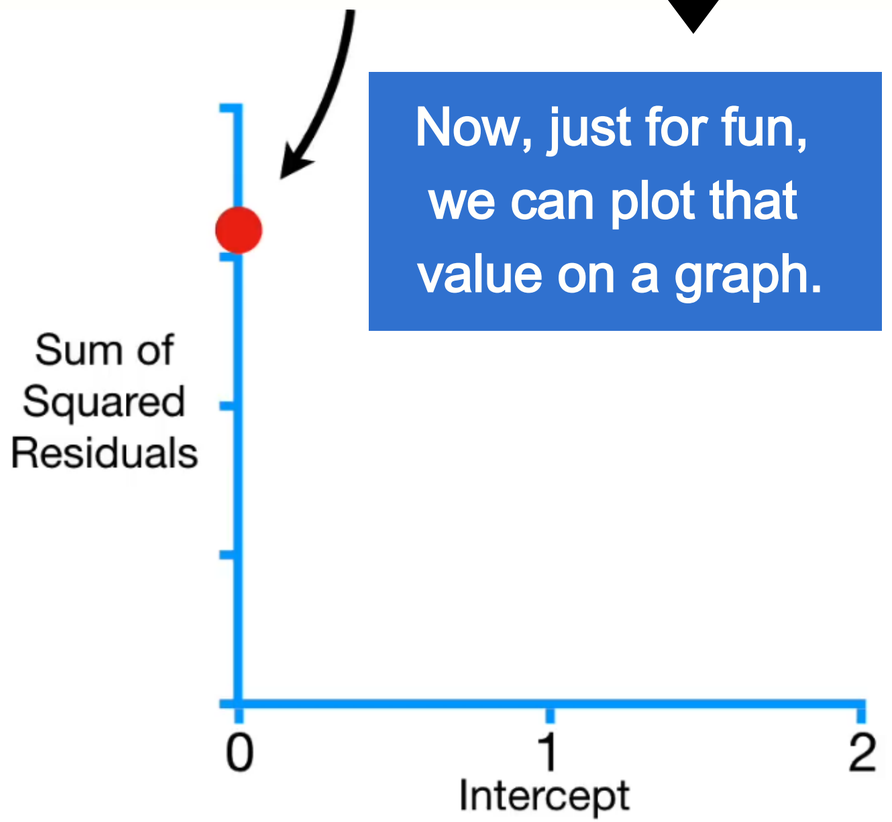
Now, just for fun, we can plot that value on a graph.



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

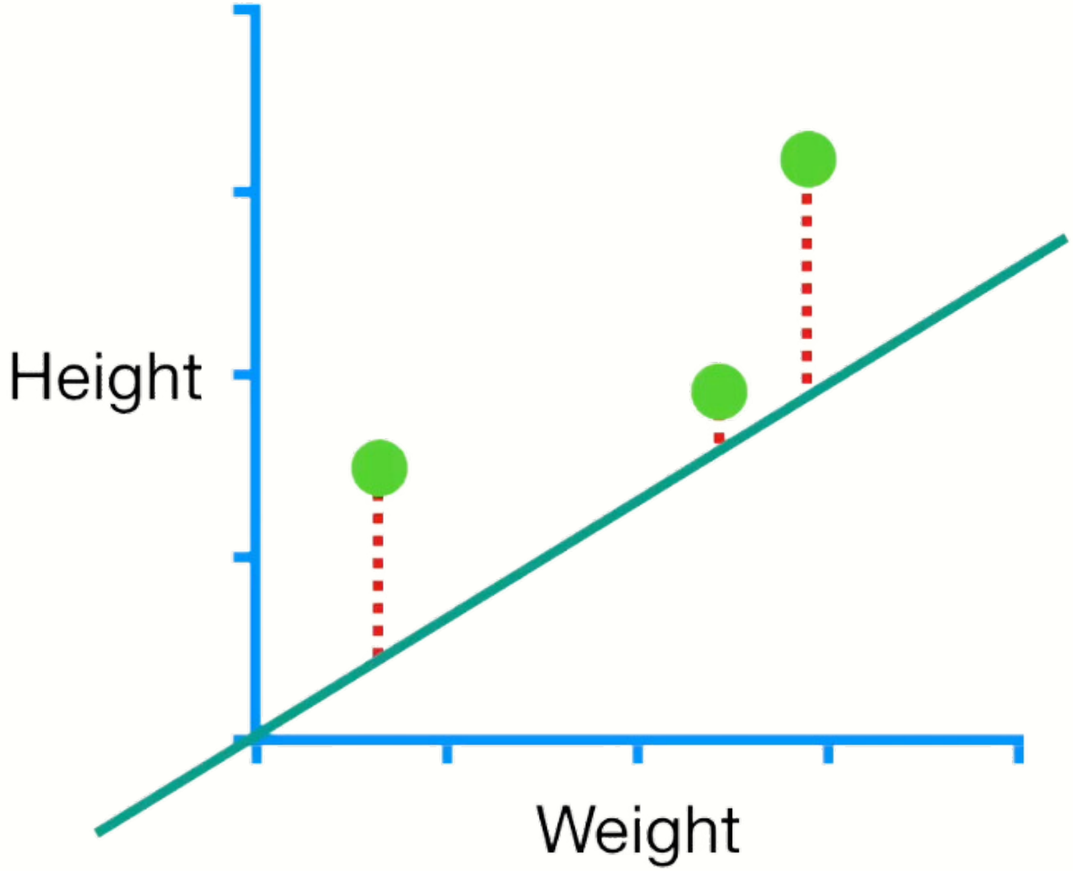


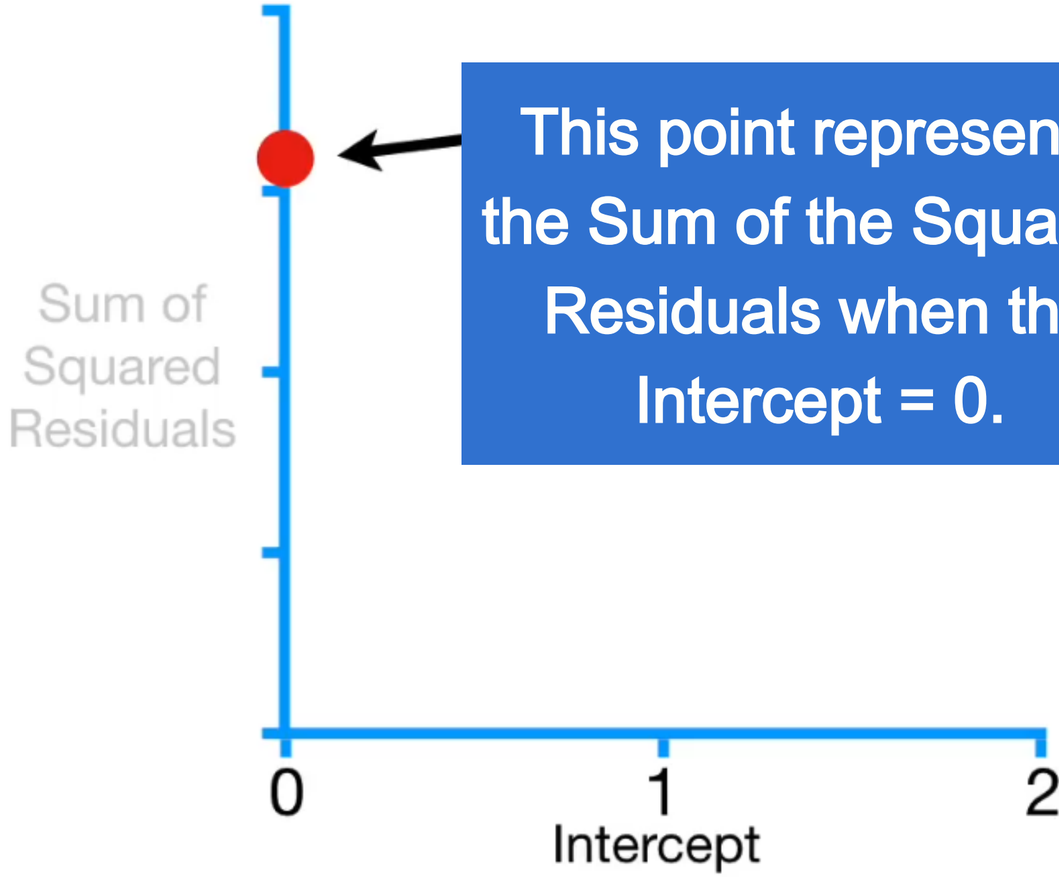
Now, just for fun, we can plot that value on a graph.



This graph has the Sum of Squared Residuals on the y -axis and different values for the Intercept on the x -axis.

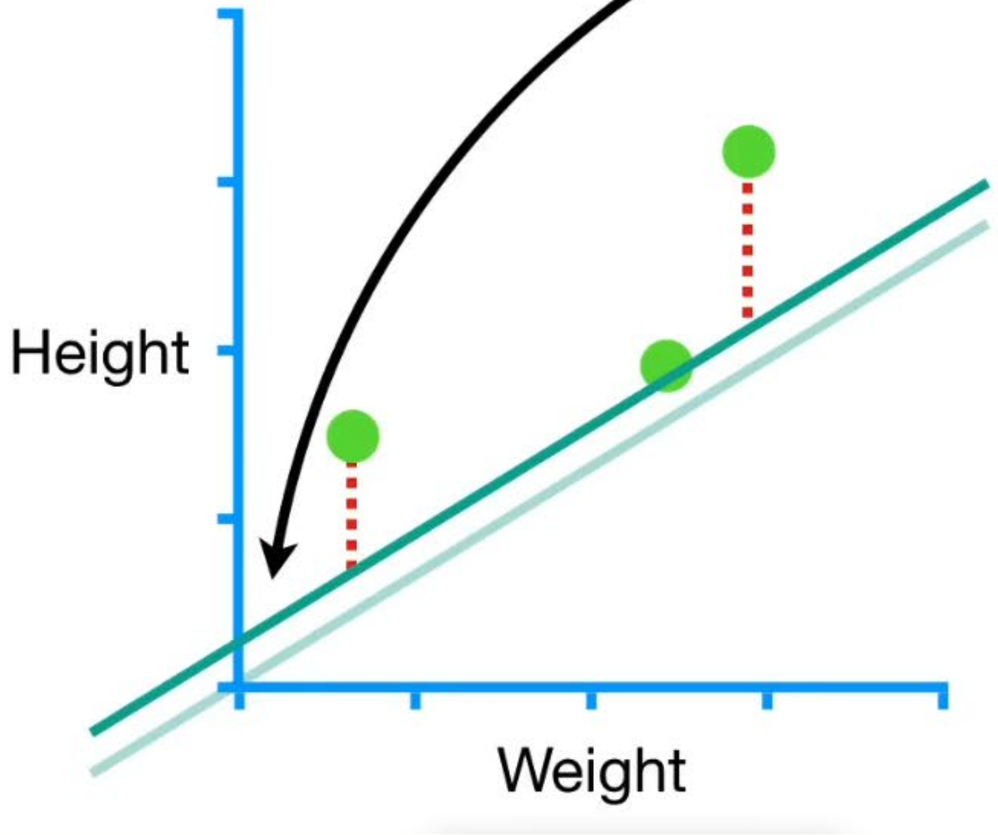
Sum of squared residuals = $1.1^2 + 0.4^2 + 1.3^2 = 3.1$

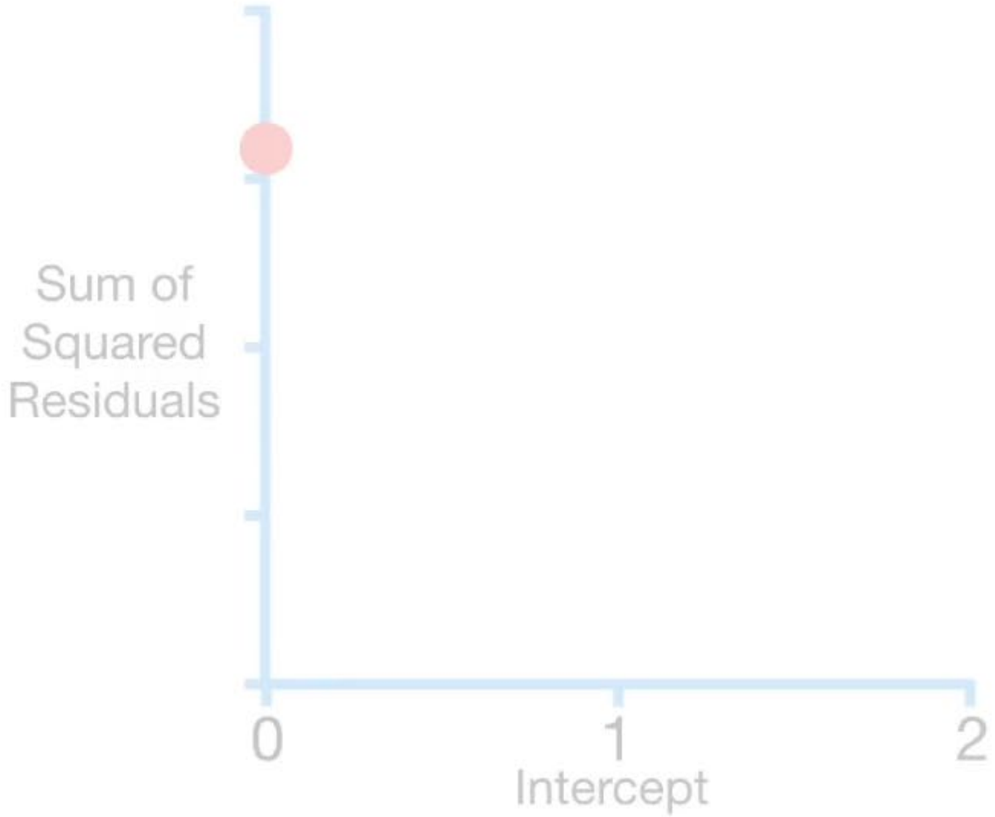




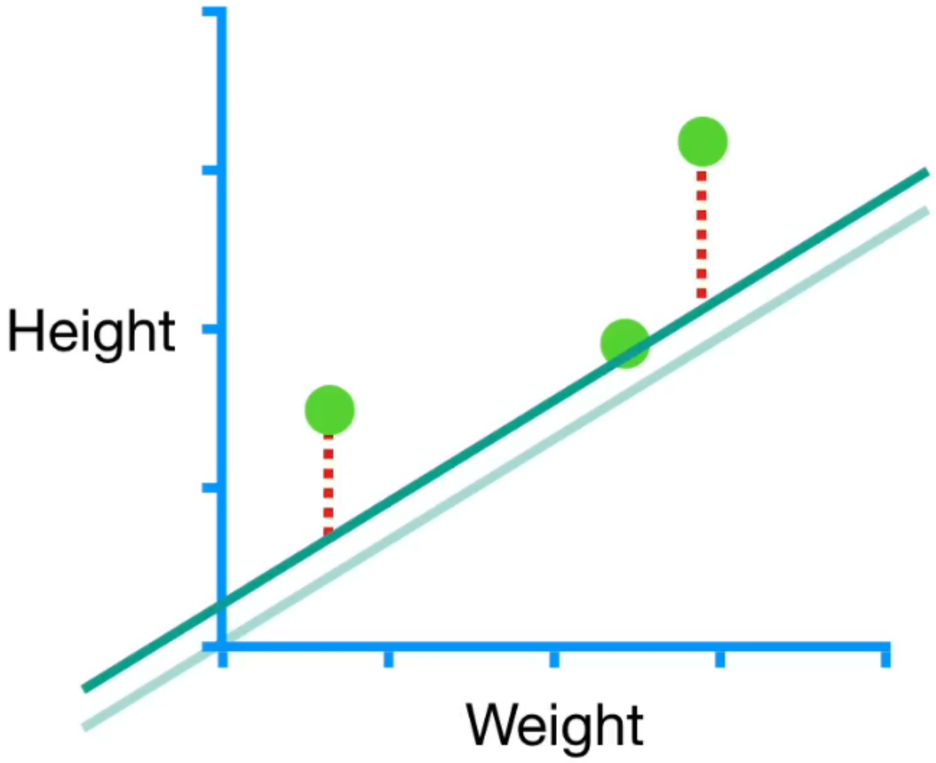
This point represents the Sum of the Squared Residuals when the Intercept = 0.

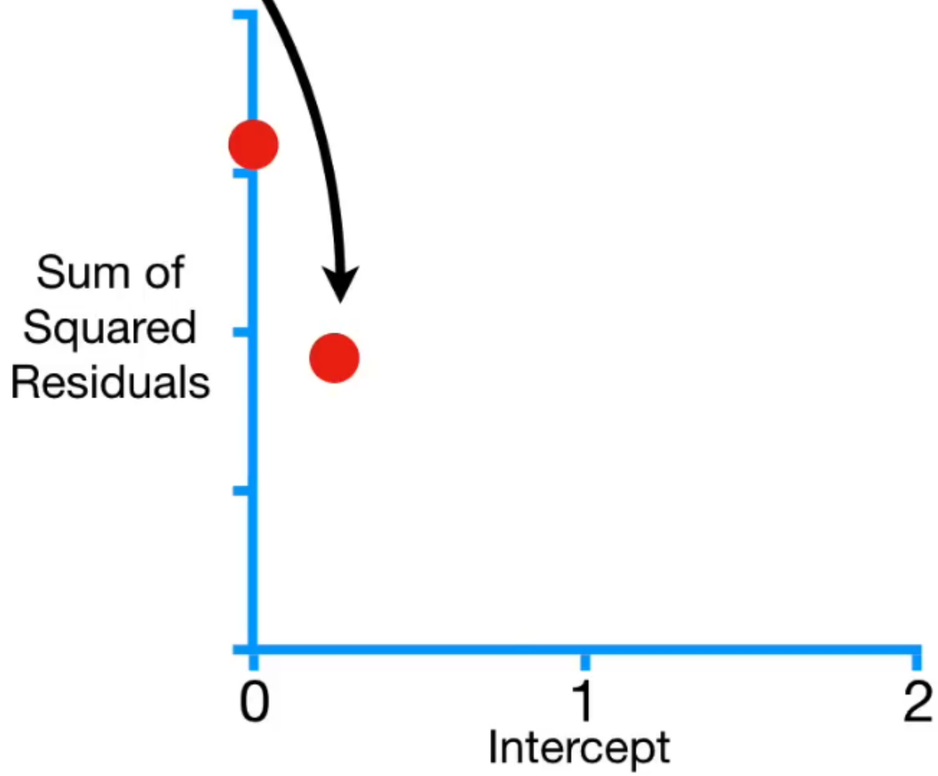
However, if the Intercept = 0.25…



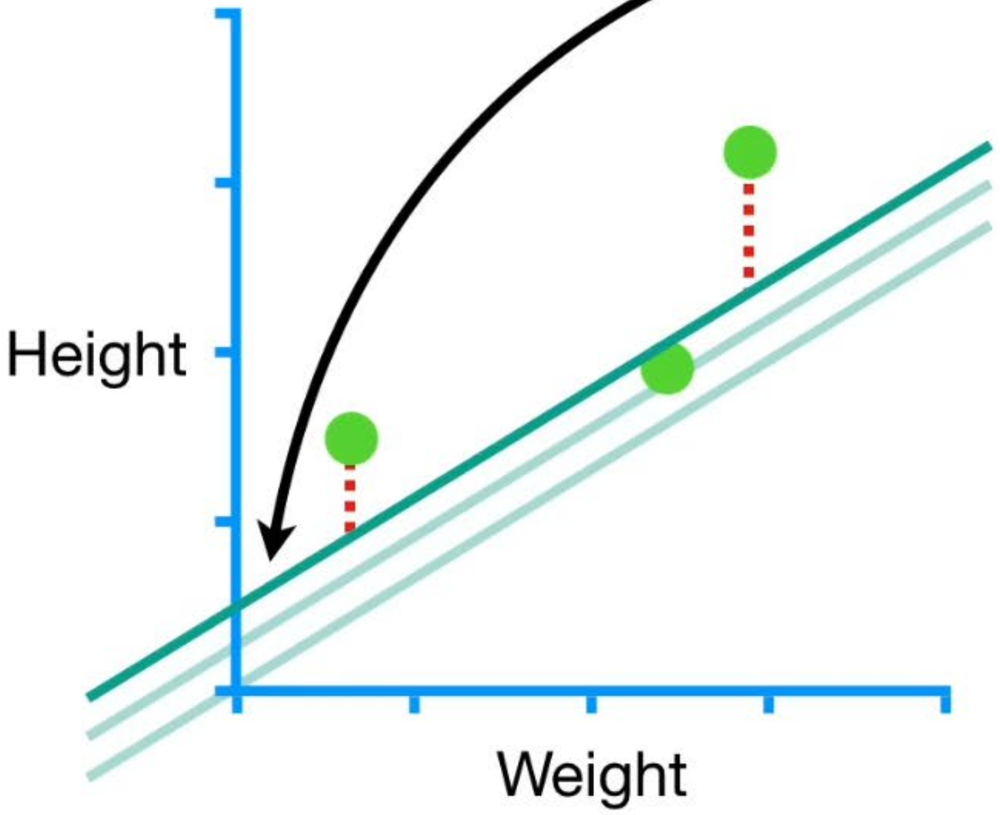


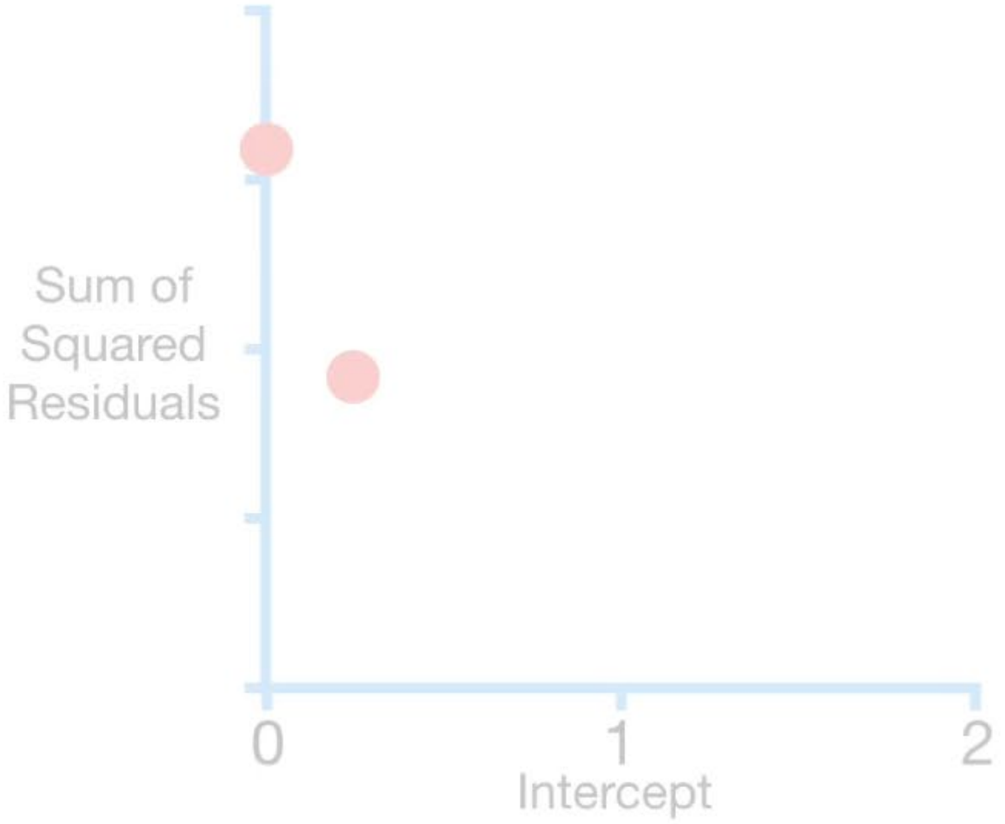
…then we would get this point on the graph.



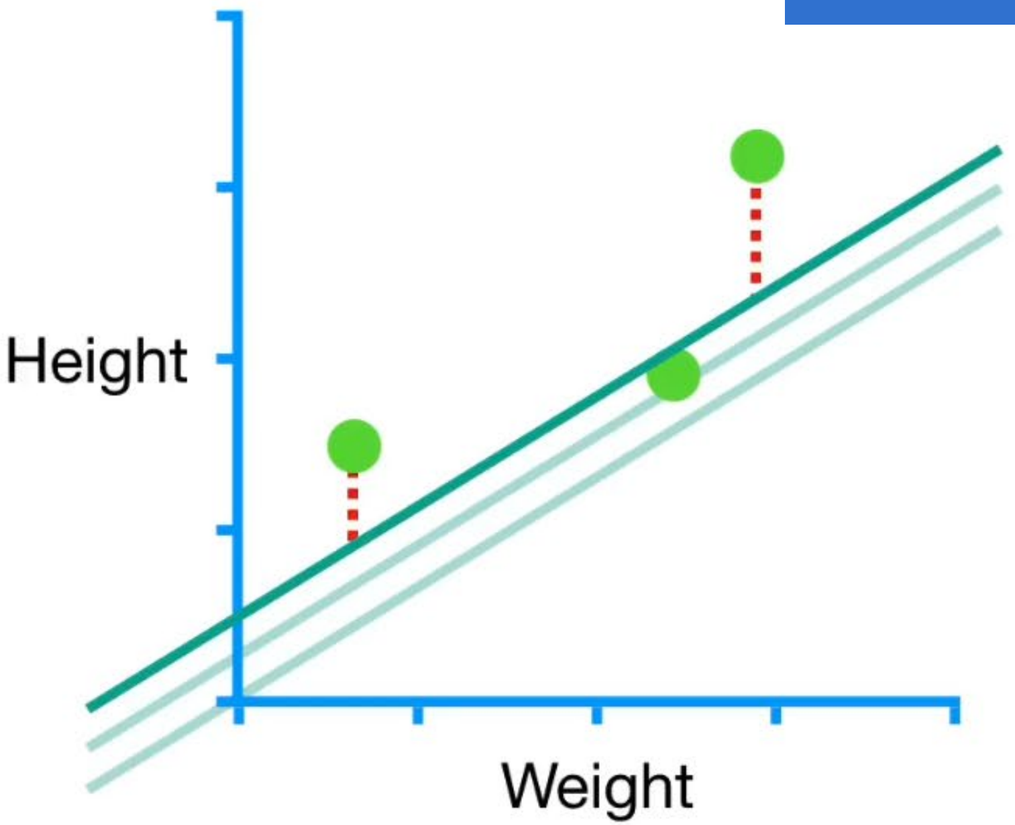


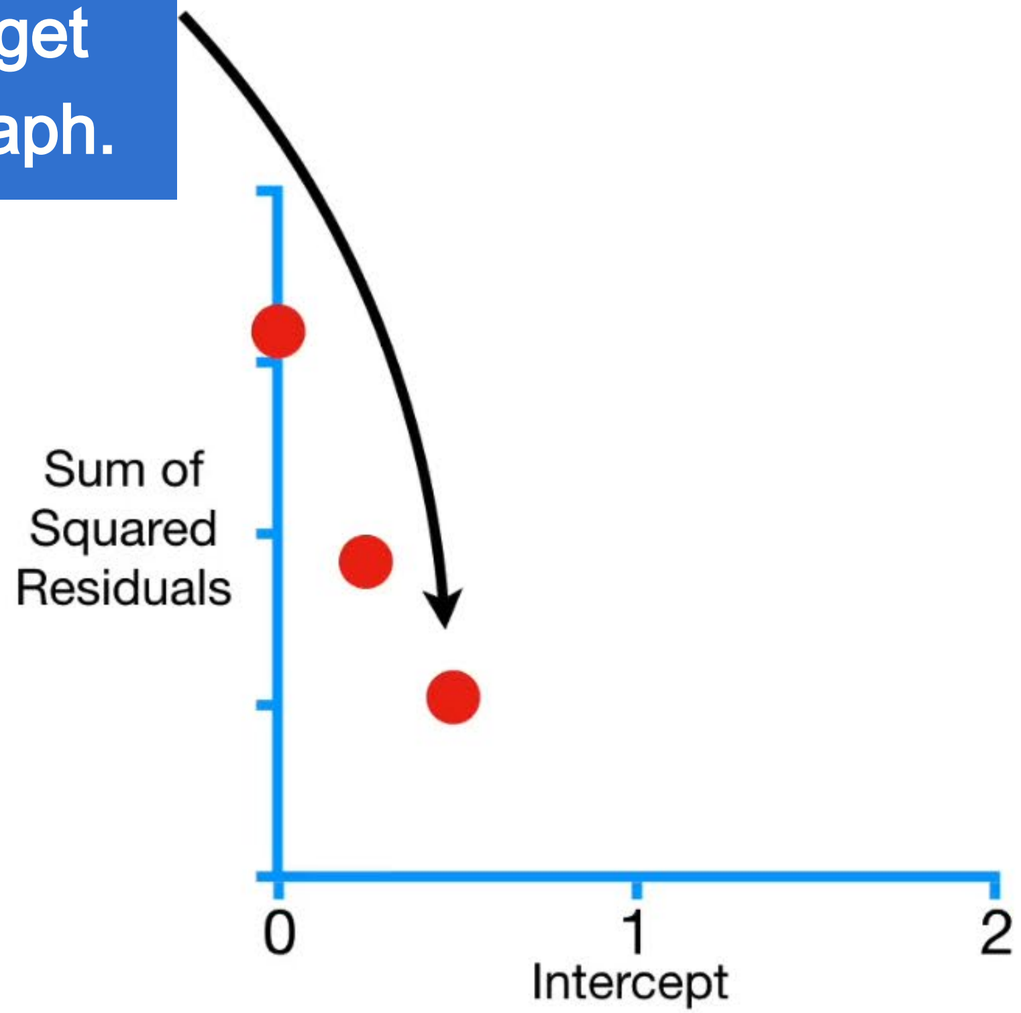
And if the Intercept = 0.5…



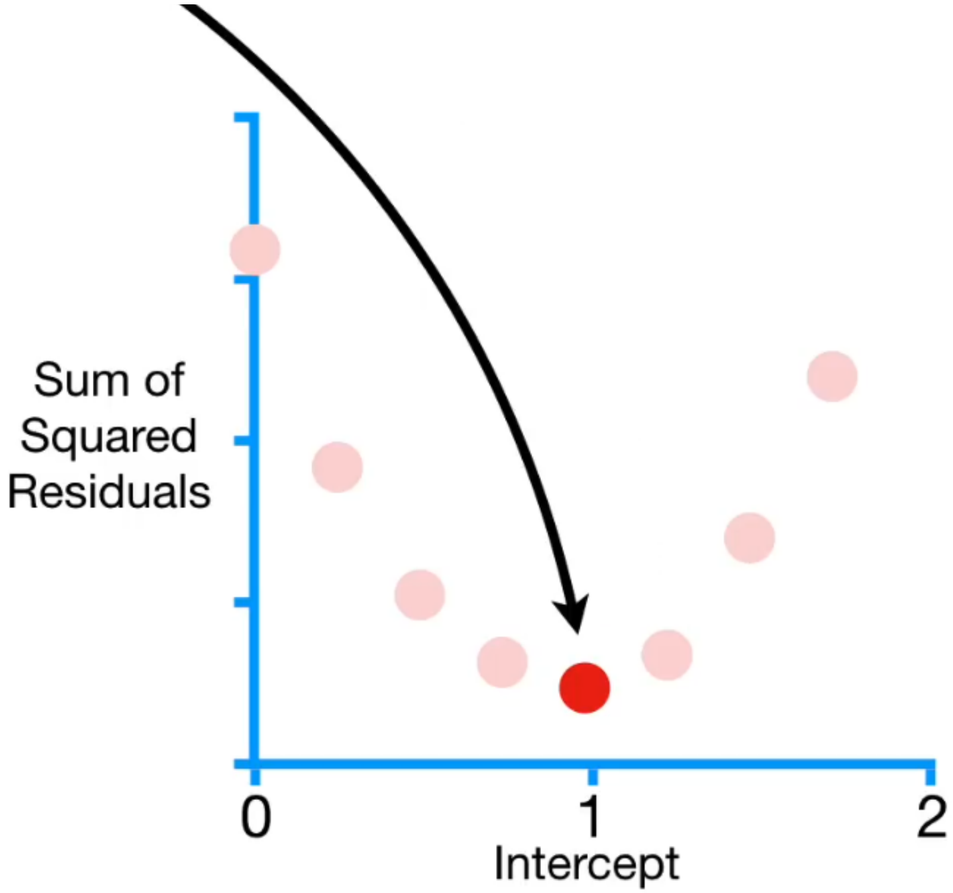


…then we would get this point on the graph.

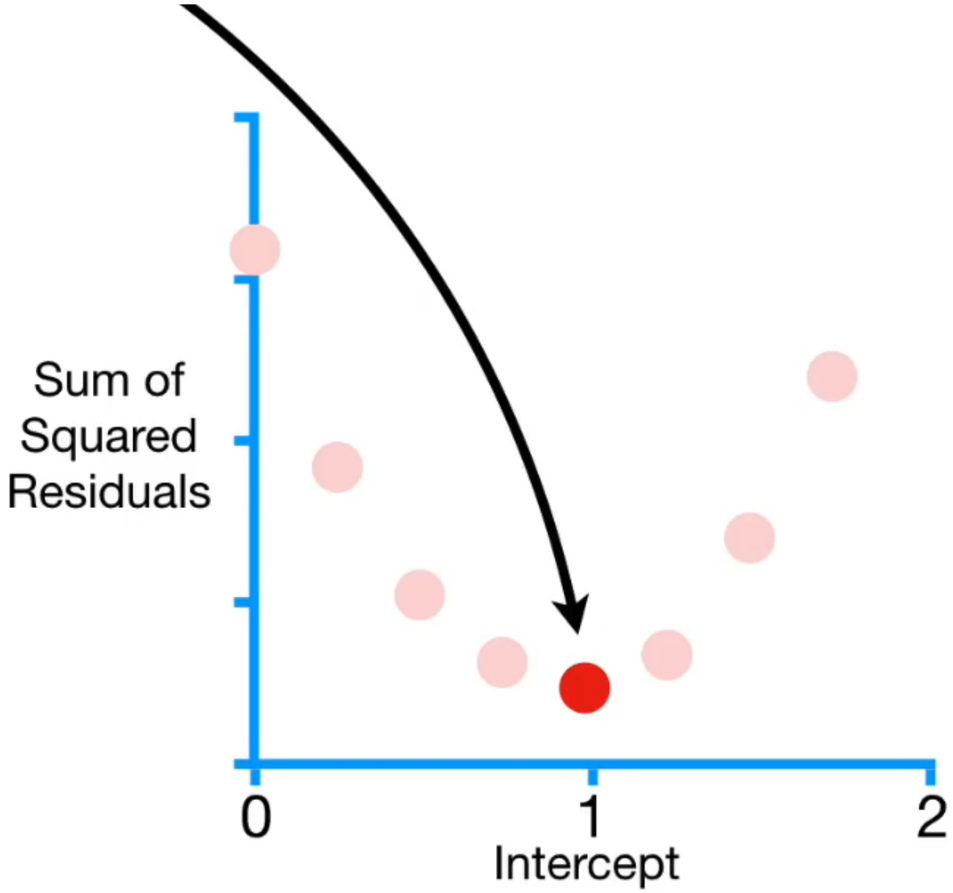




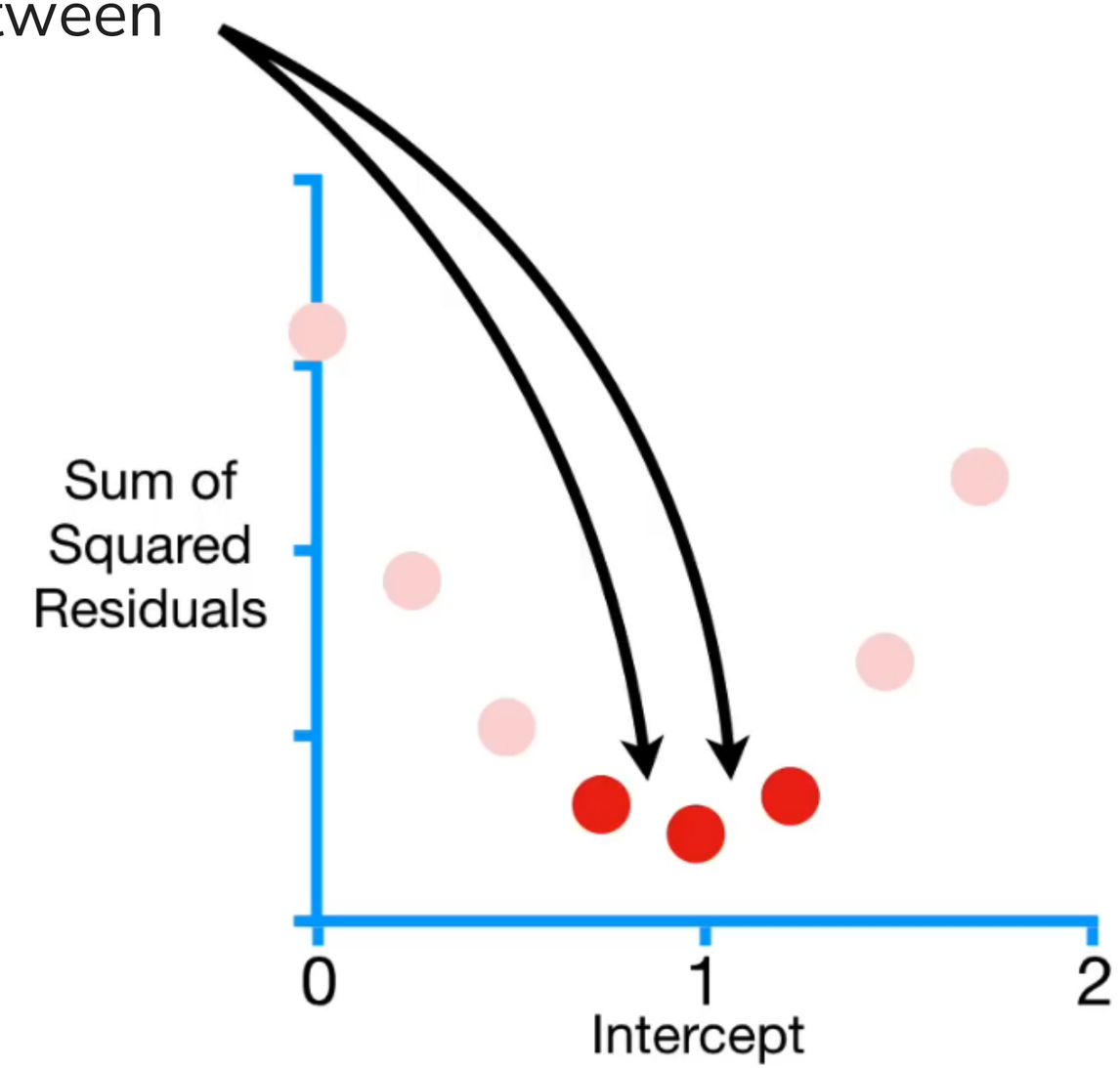
Of the points that we calculated for the graph, this one has the lowest Sum of Squared Residuals...



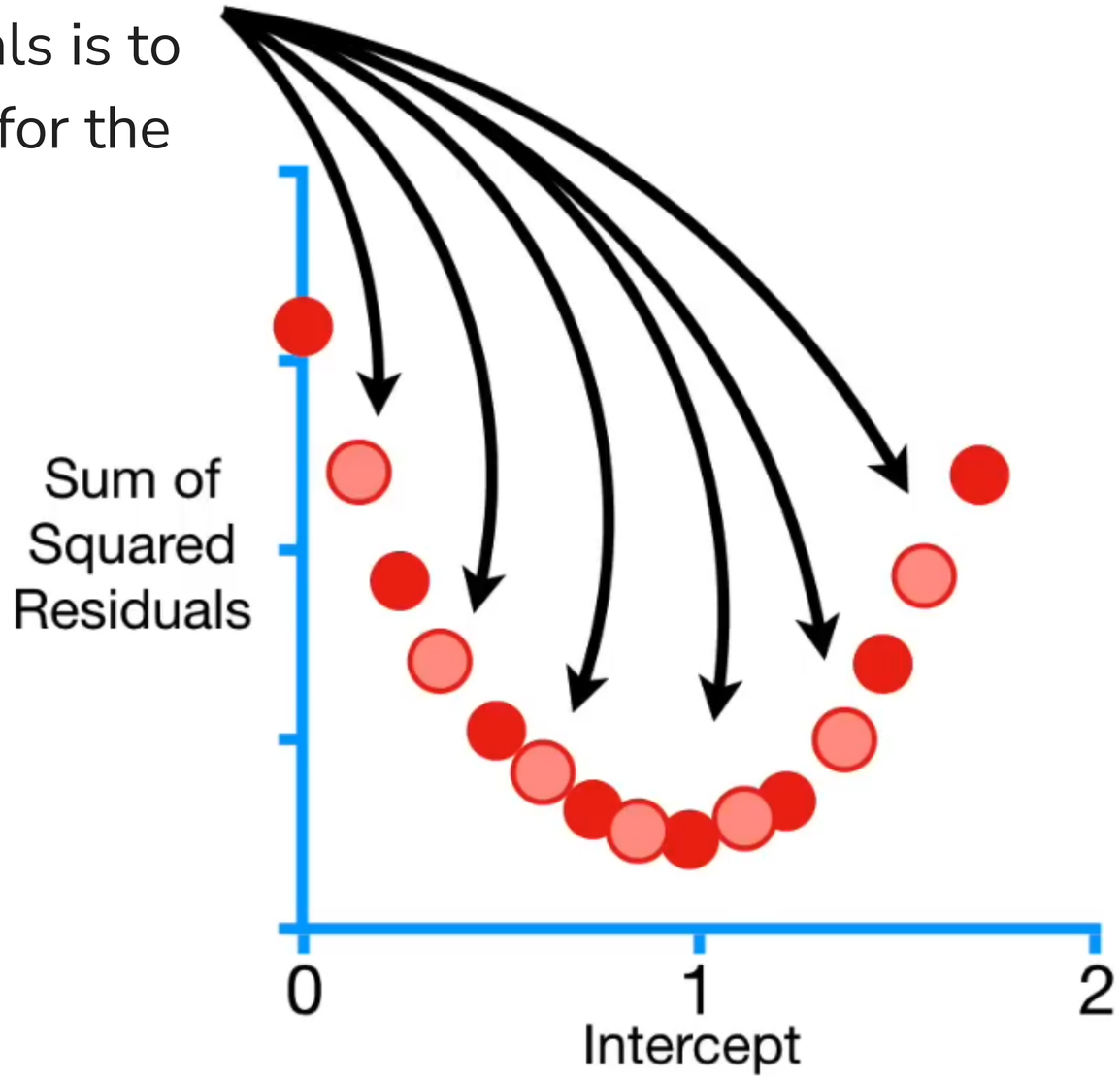
...but is it the best we can do?



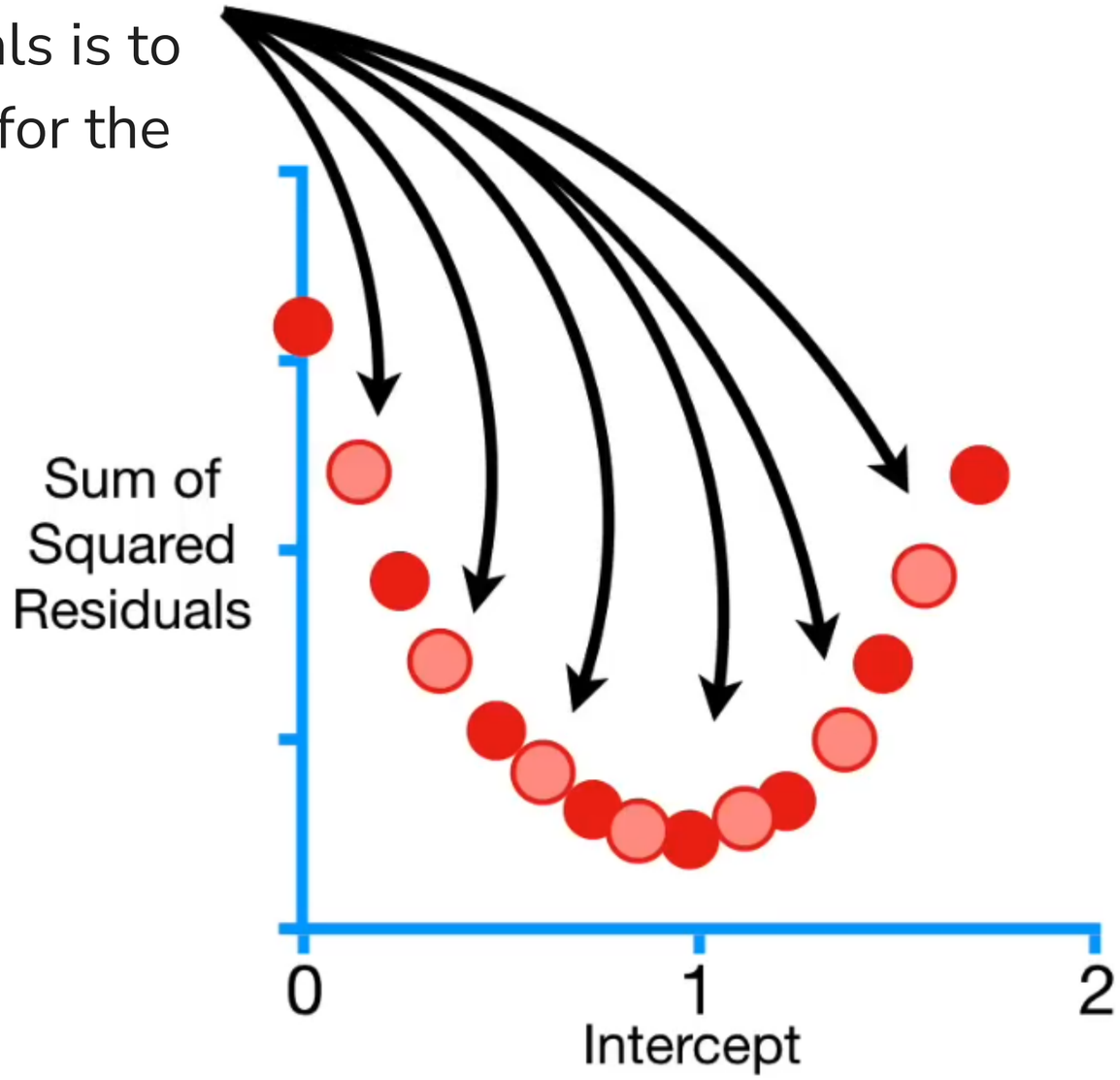
What if the best value for the Intercept is somewhere between these values?



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** .

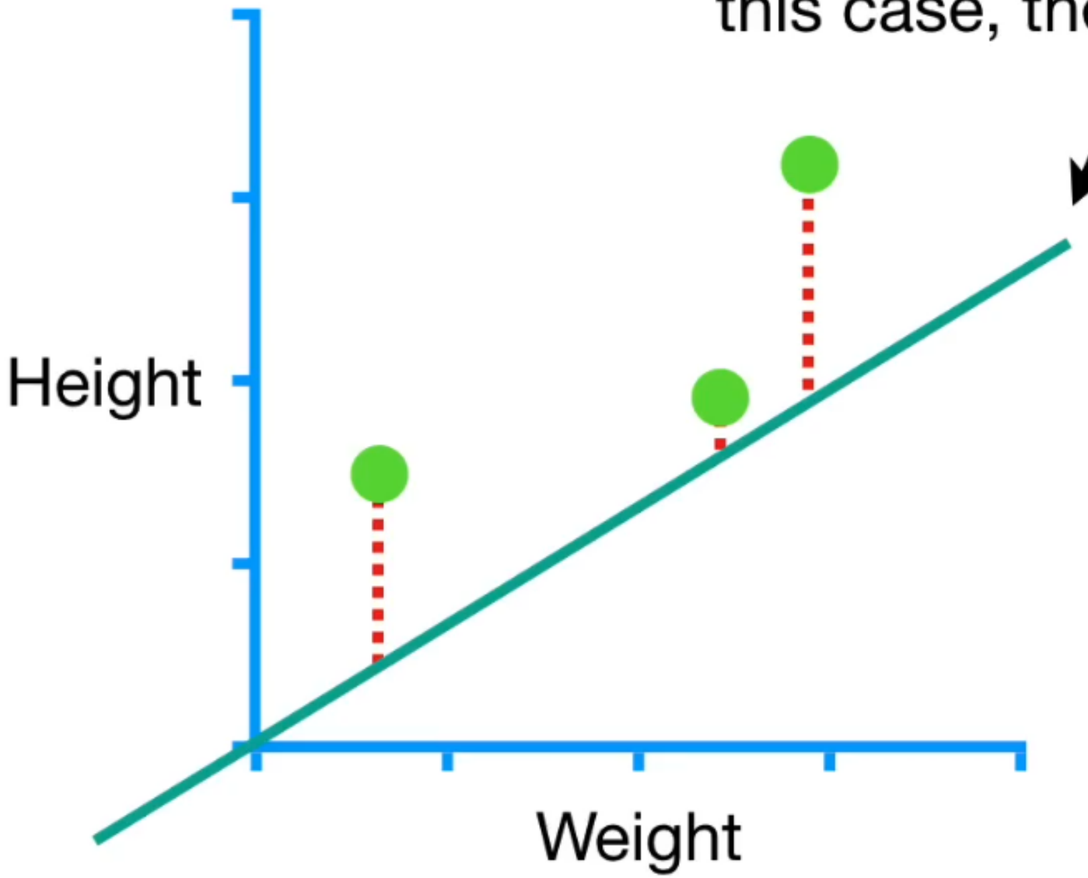


A slow and painful method for finding the minimal Sum of the Squared Residuals is to plug and chug a bunch more values for the **Intercept** .



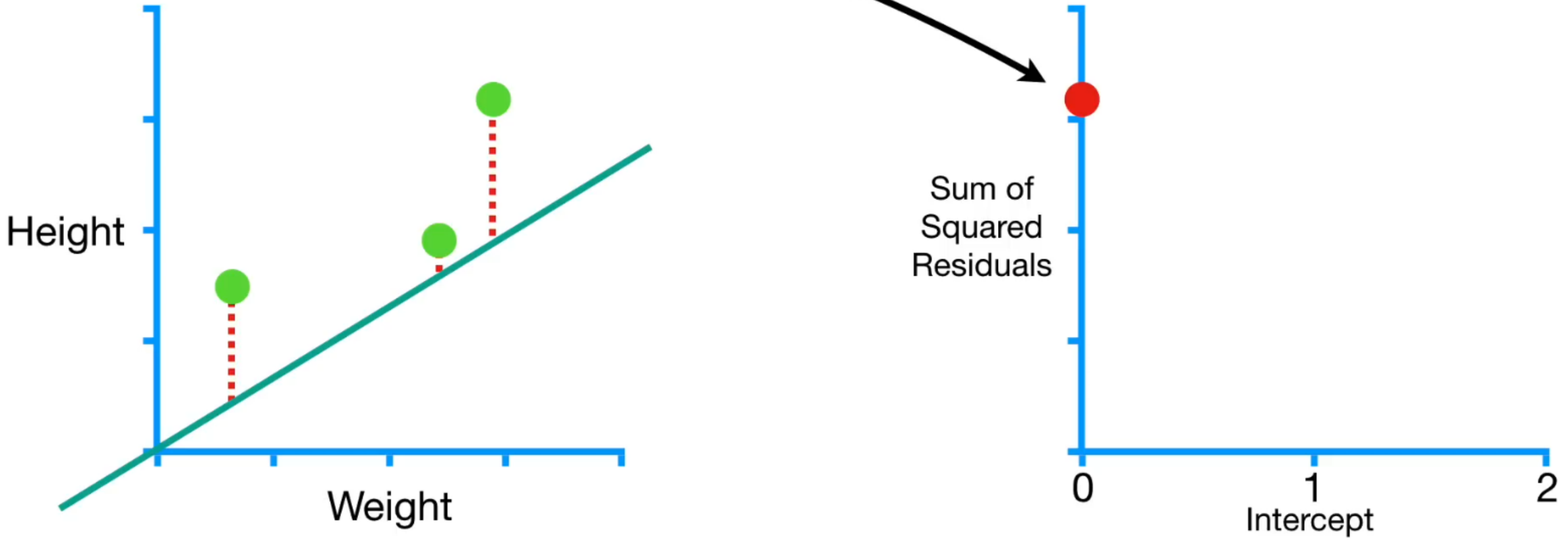
Don’t despair! Gradient Descent is way more efficient!

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**.

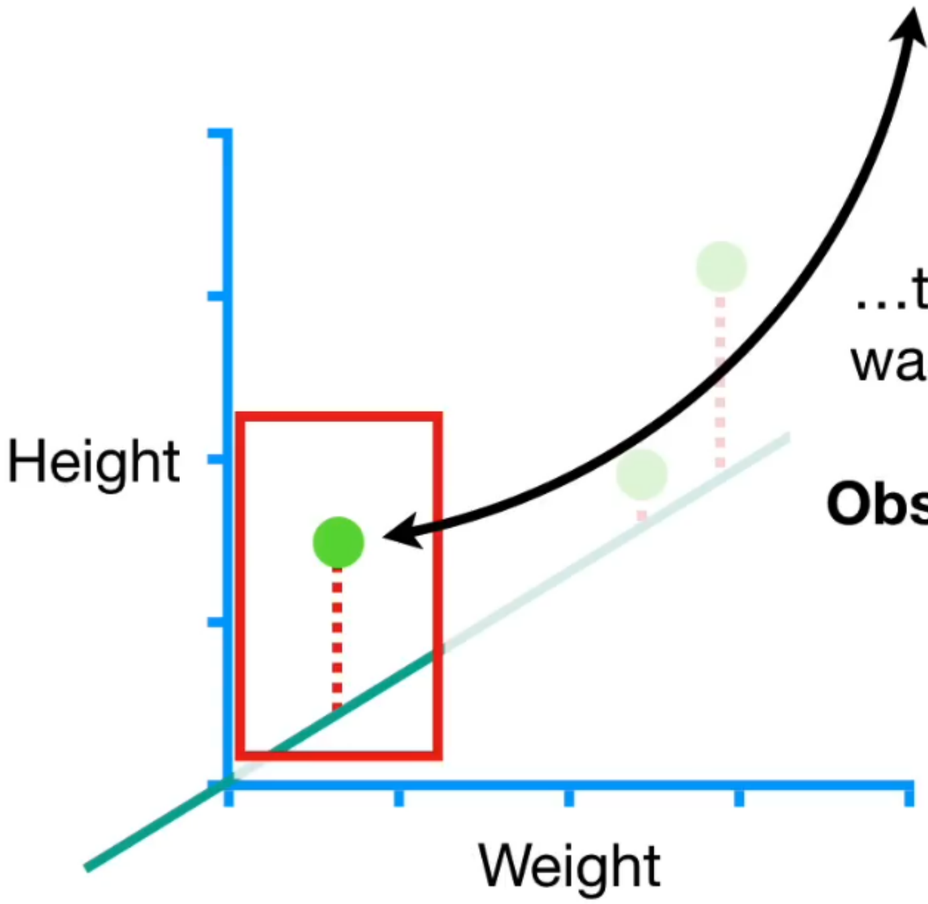


Sum of squared residuals

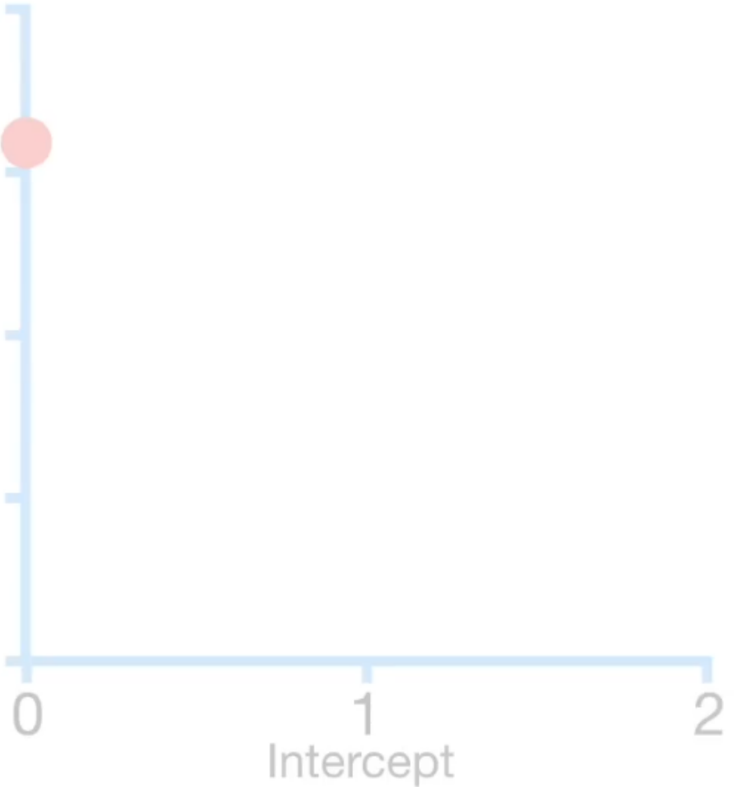
When we calculated the Sum of the Squared Residuals…



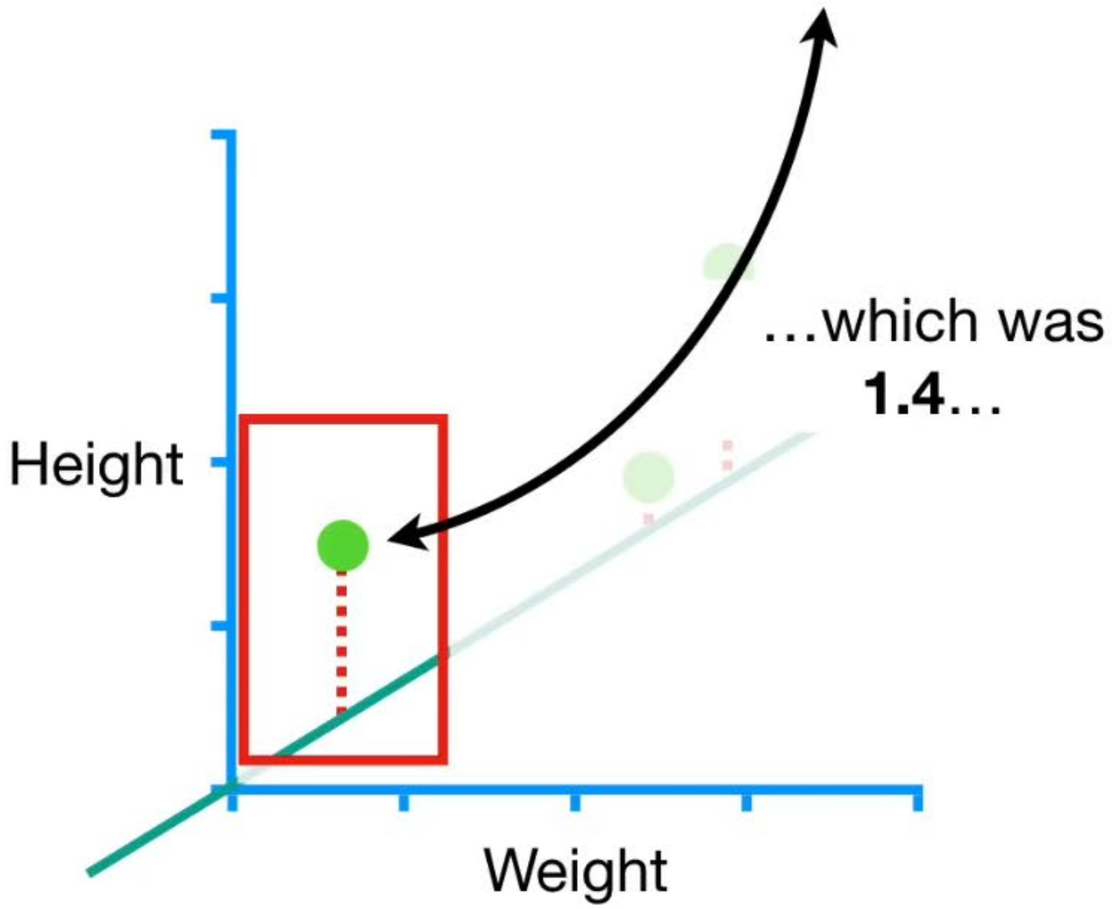
Sum of squared residuals = (observed - predicted)^2

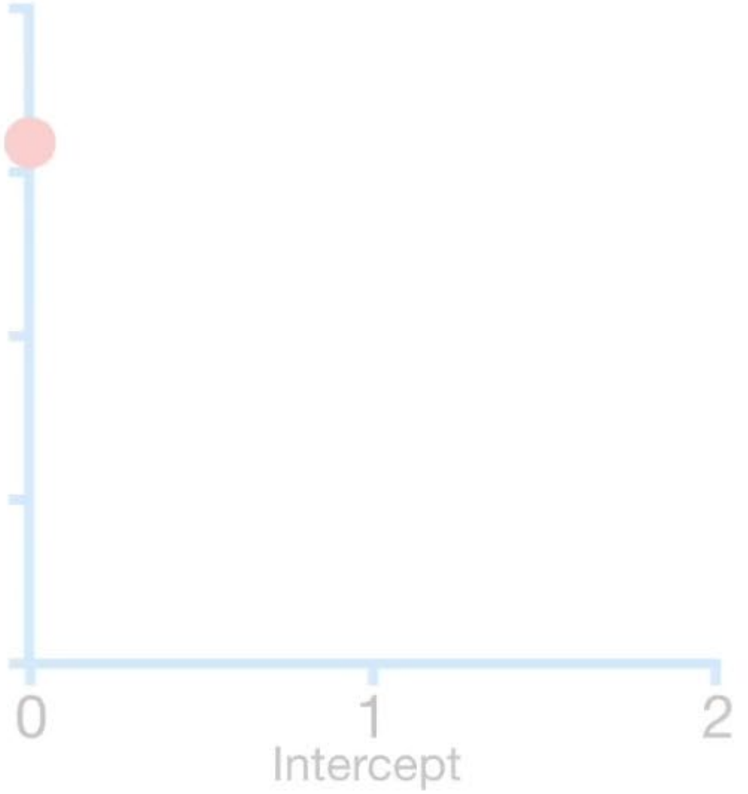


…the first residual was the difference between the **Observed Height**…

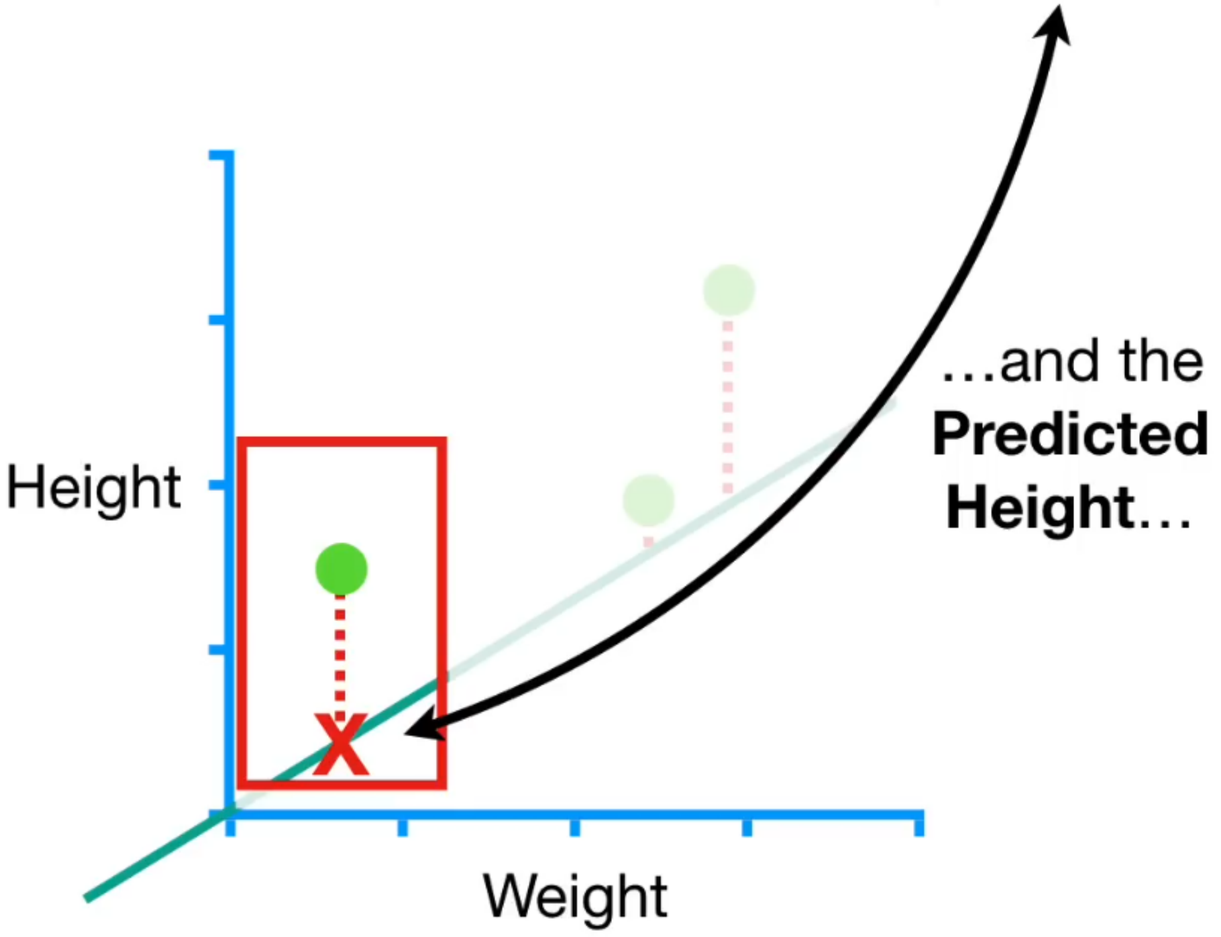


Sum of squared residuals = (**1.4** - predicted)^2

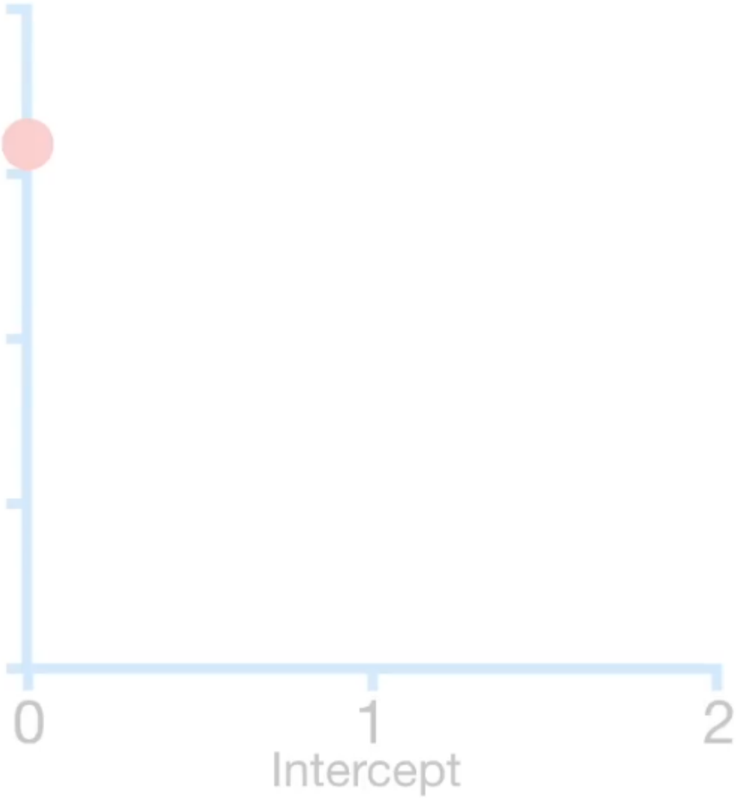




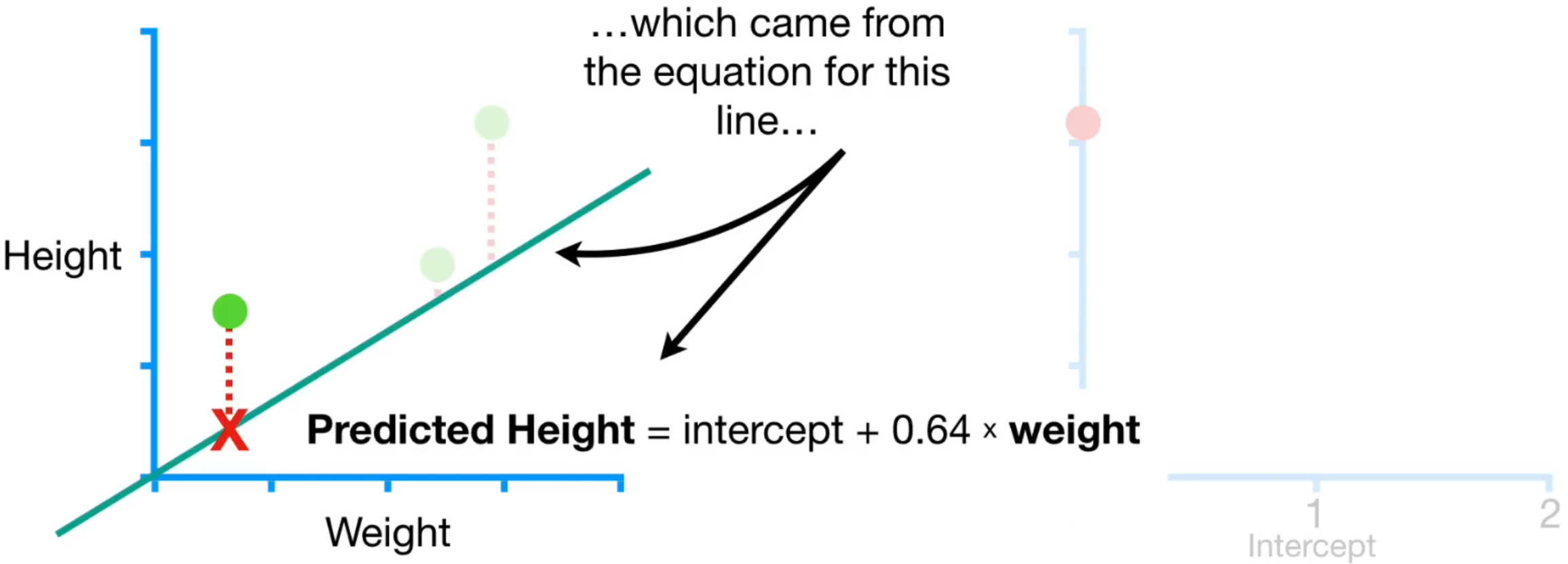
Sum of squared residuals = (**1.4** - predicted)^2



…and the **Predicted Height**…

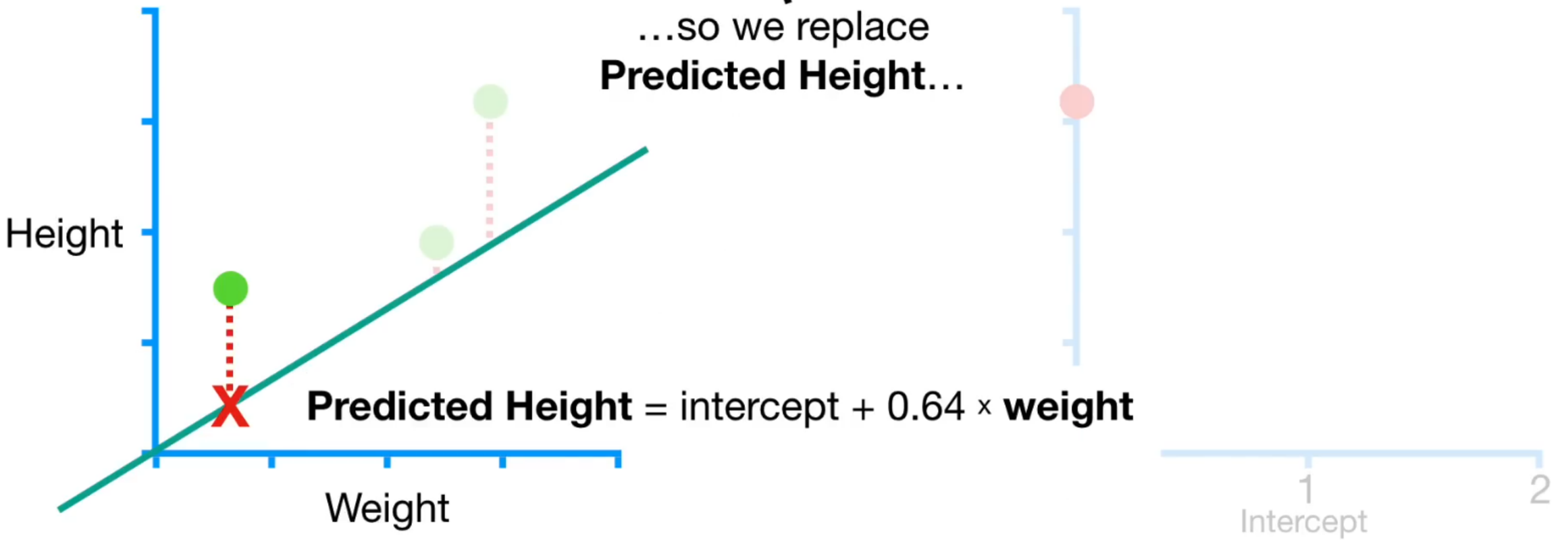


Sum of squared residuals = (**1.4** - predicted)^2

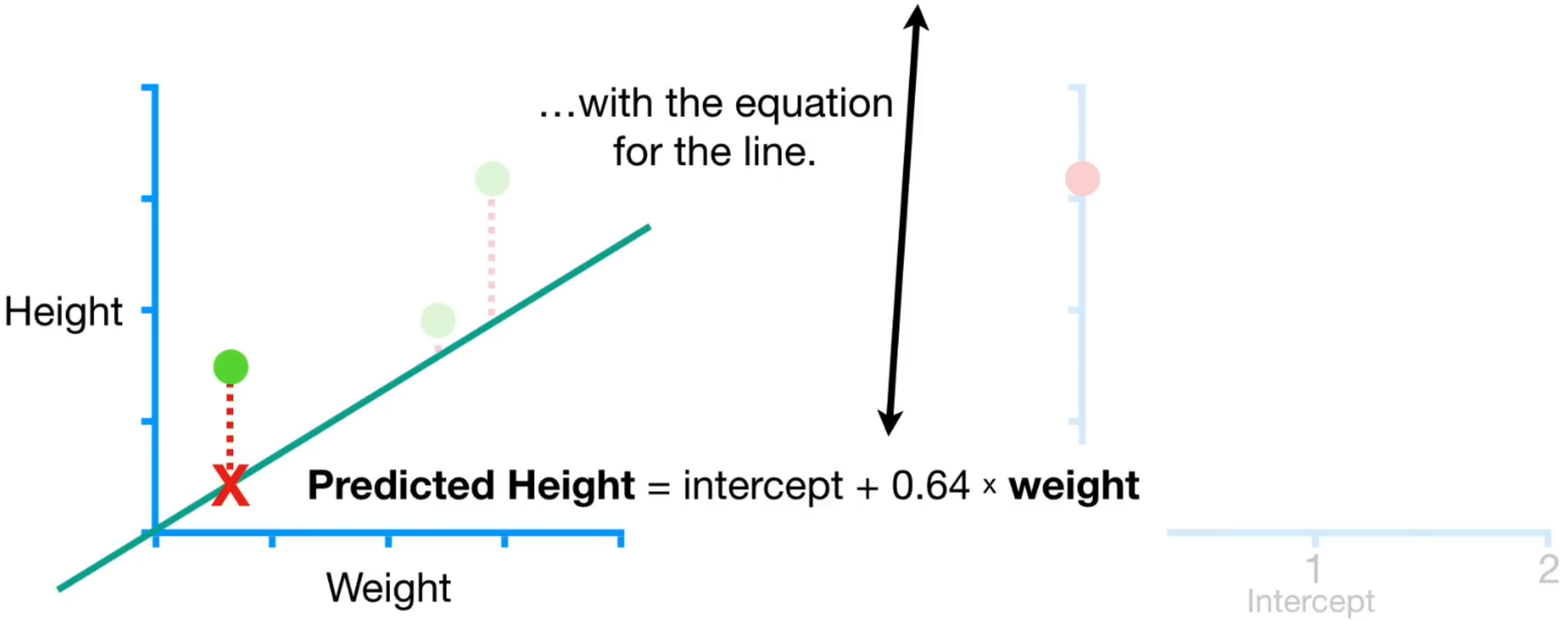


Sum of squared residuals = (**1.4** - predicted)^2

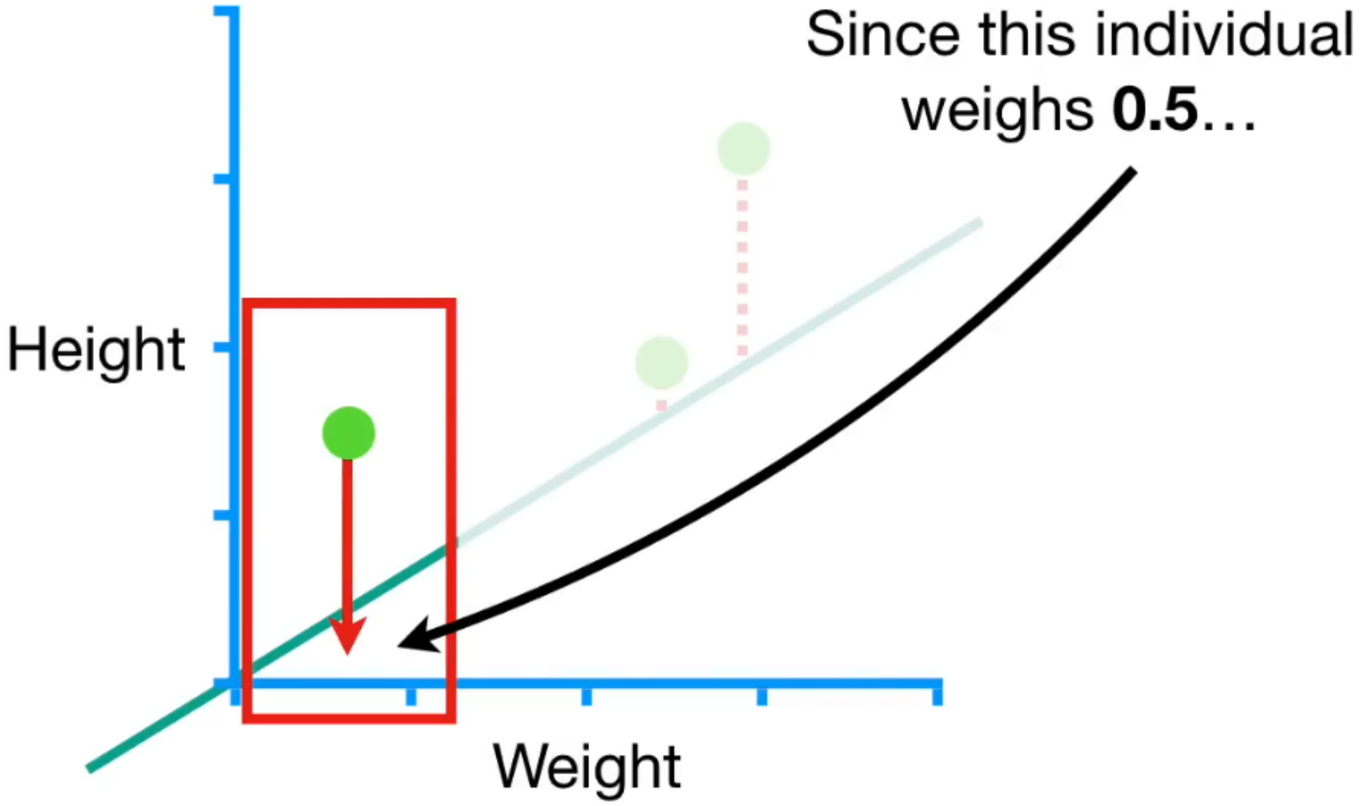
…so we replace **Predicted Height**…

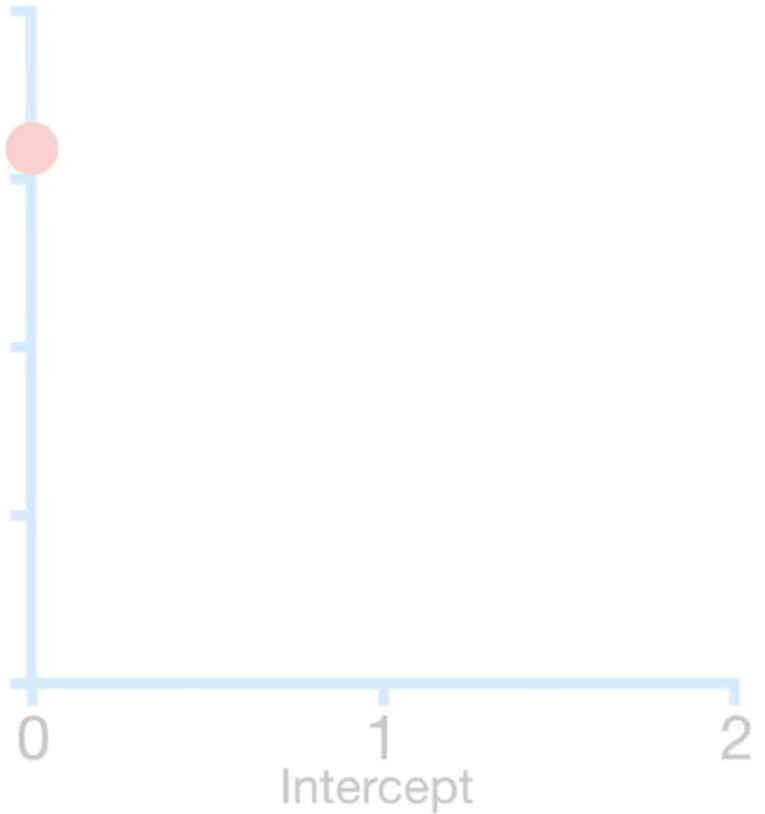


Sum of squared residuals = (**1.4** - (intercept + 0.64 × **weight**))^2



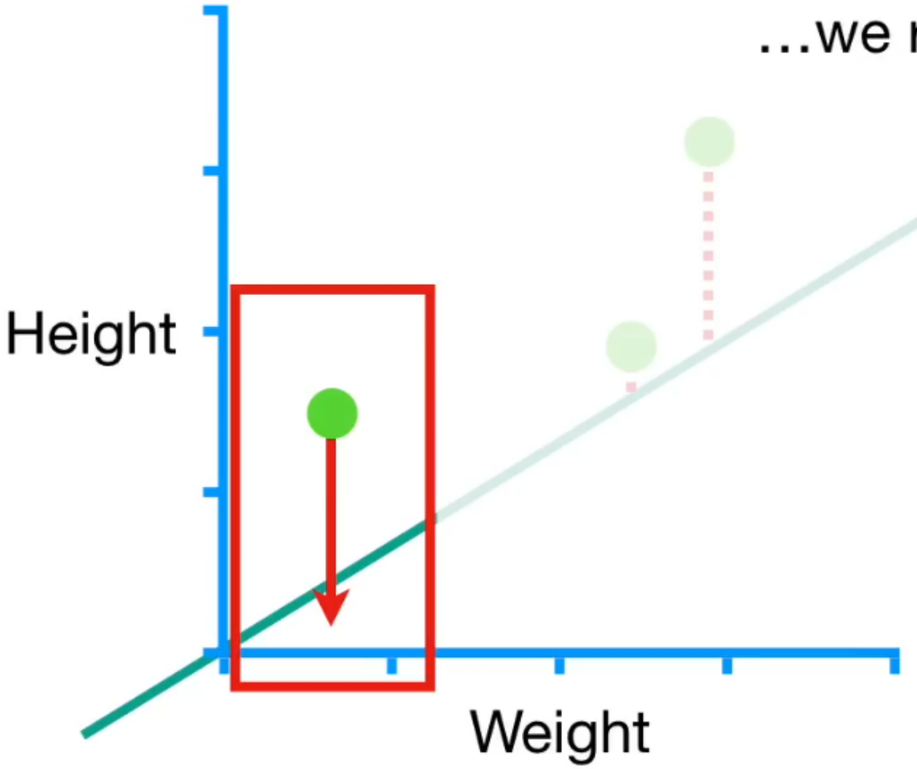
Sum of squared residuals = (**1.4** - (intercept + 0.64 × **weight**))^2

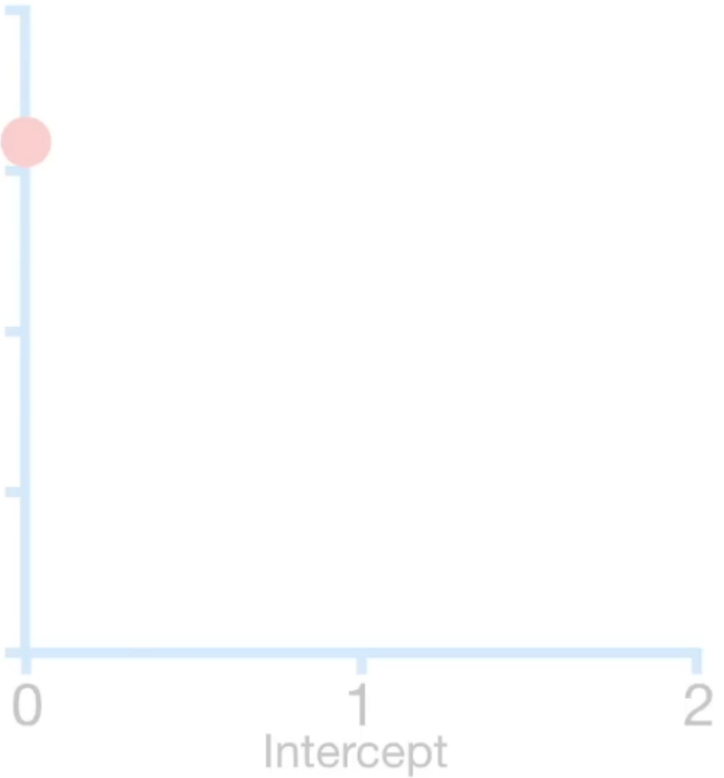




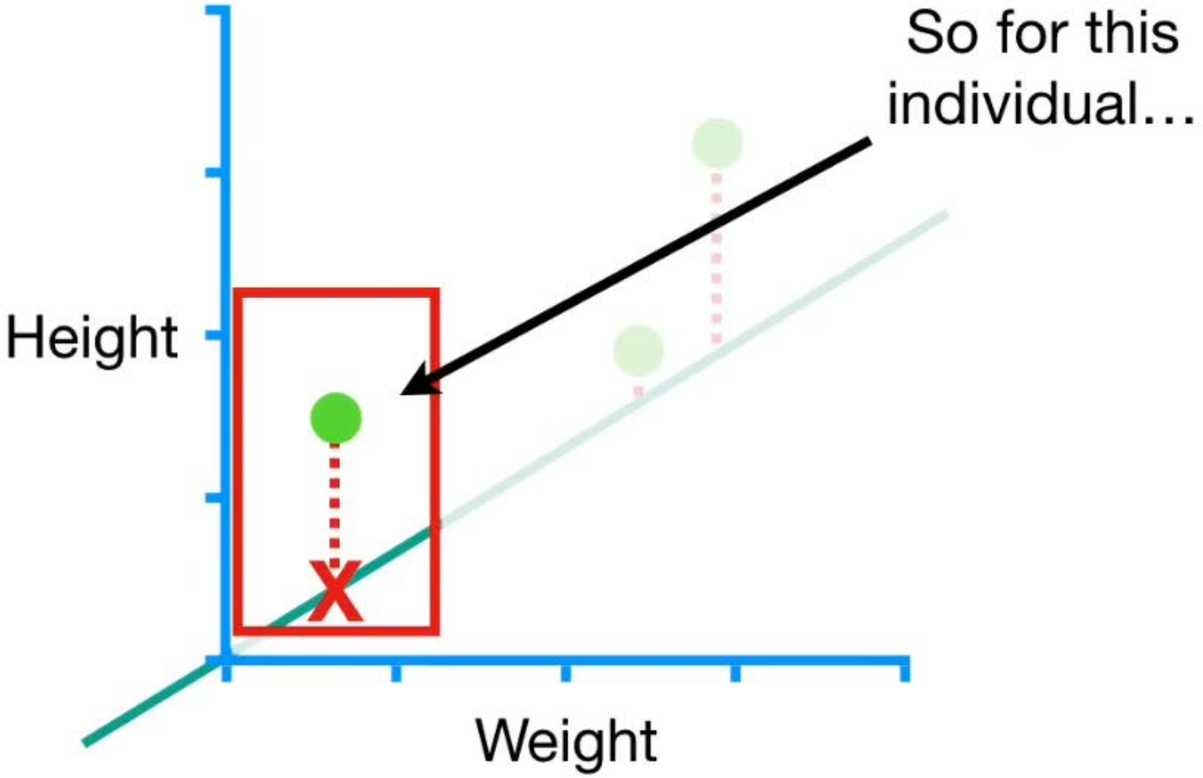
Sum of squared residuals = (**1.4** - (intercept + 0.64 × **weight**))²

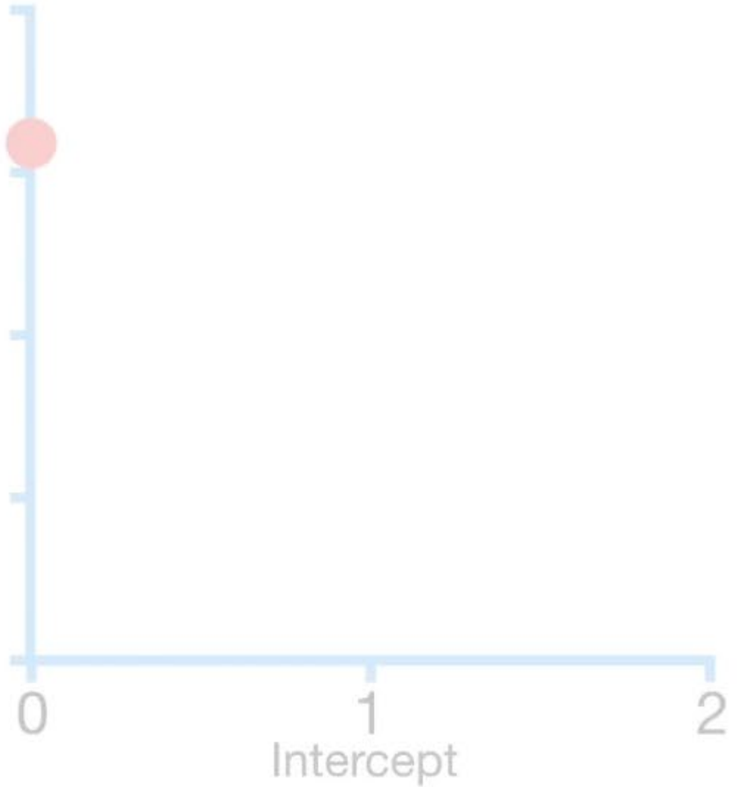
…we replace **weight** with **0.5**.



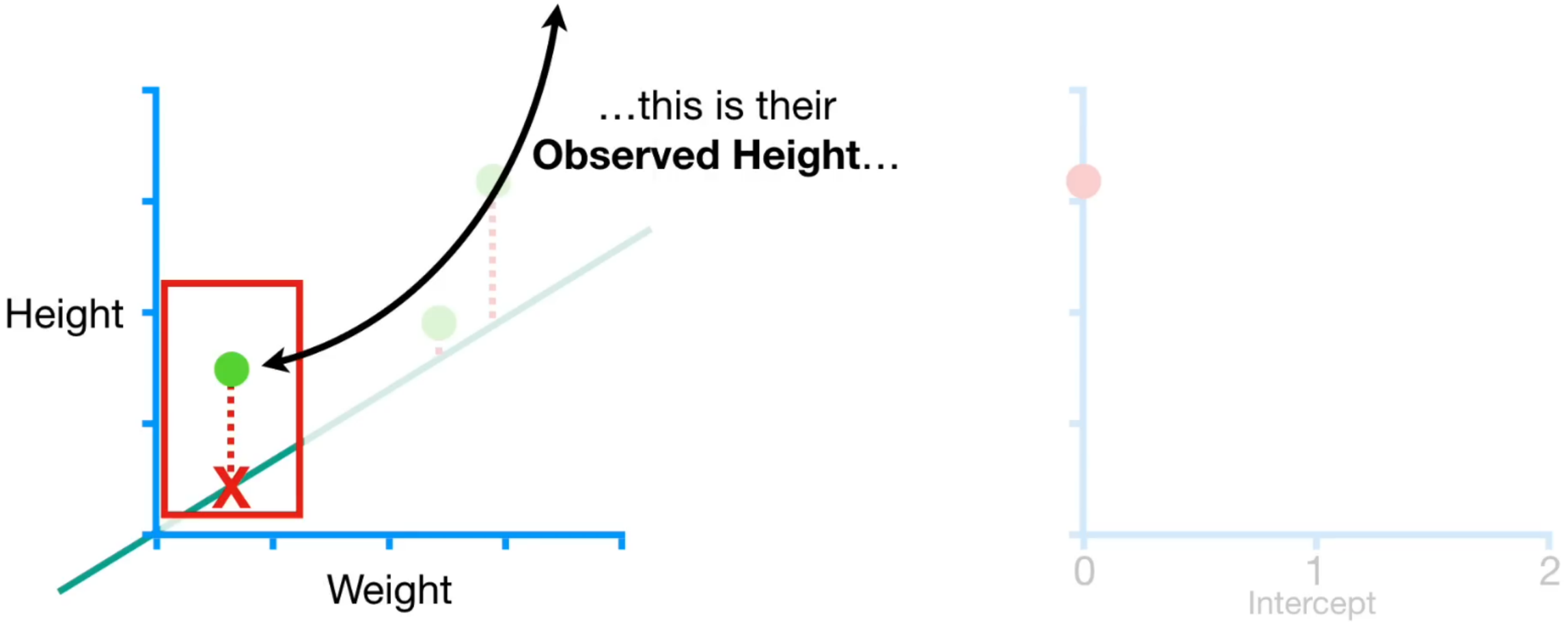


Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

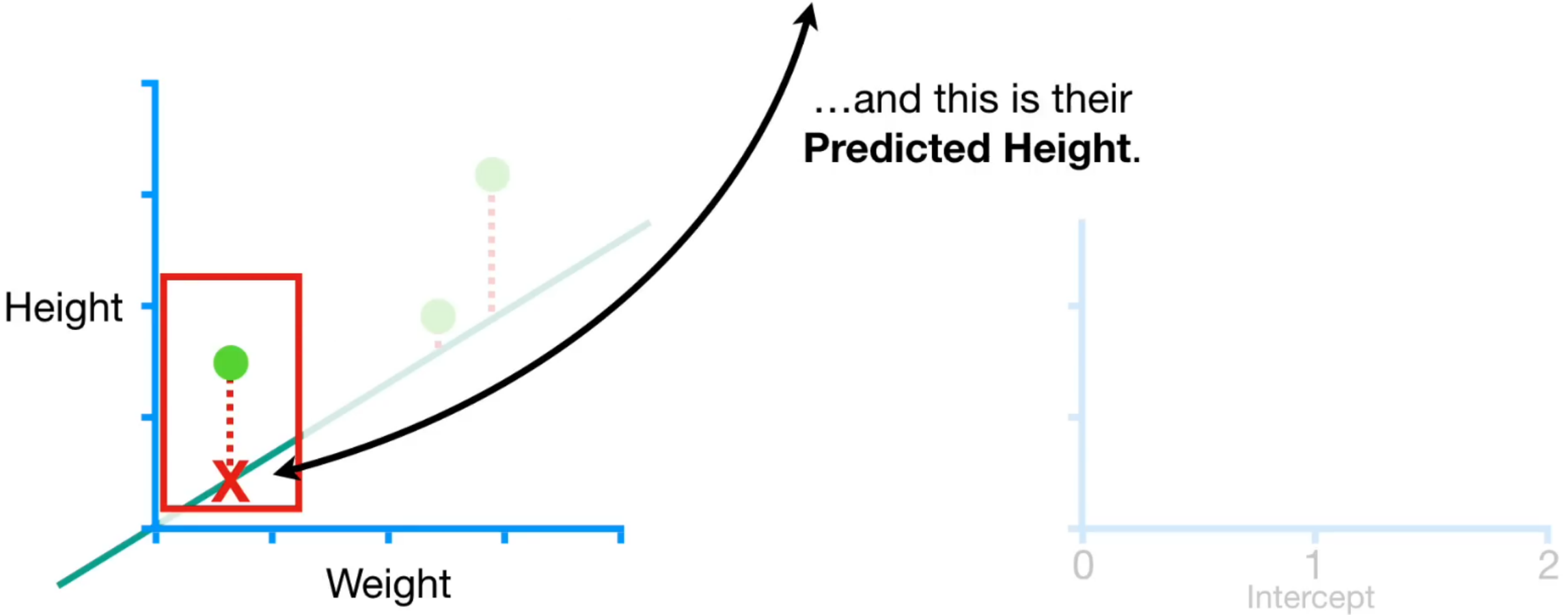




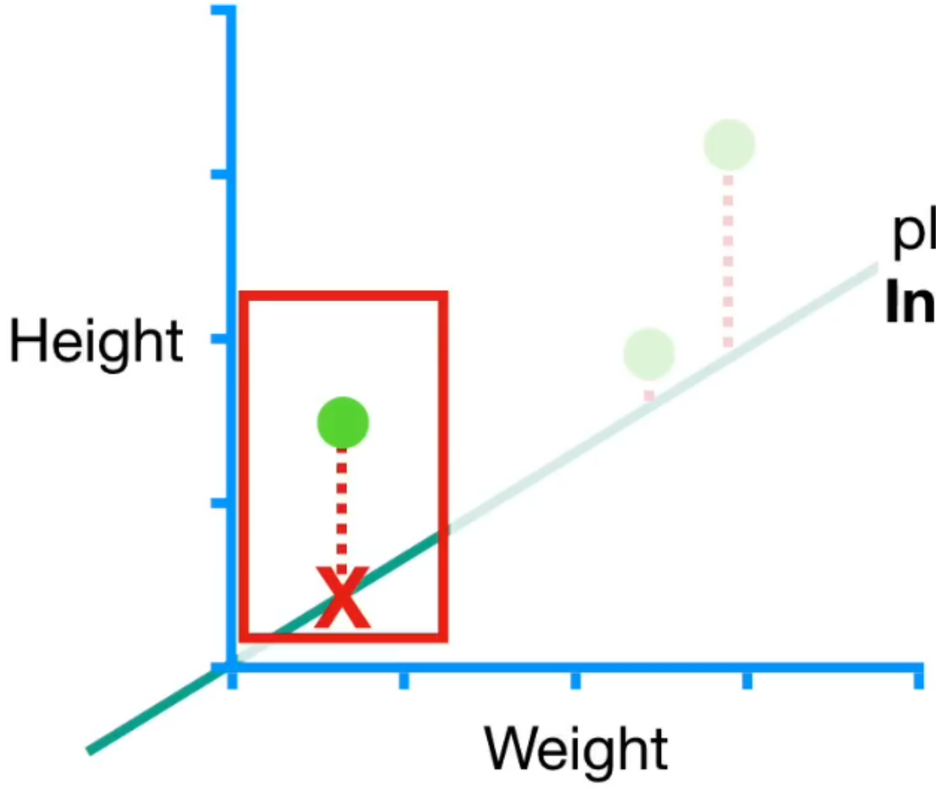
Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2



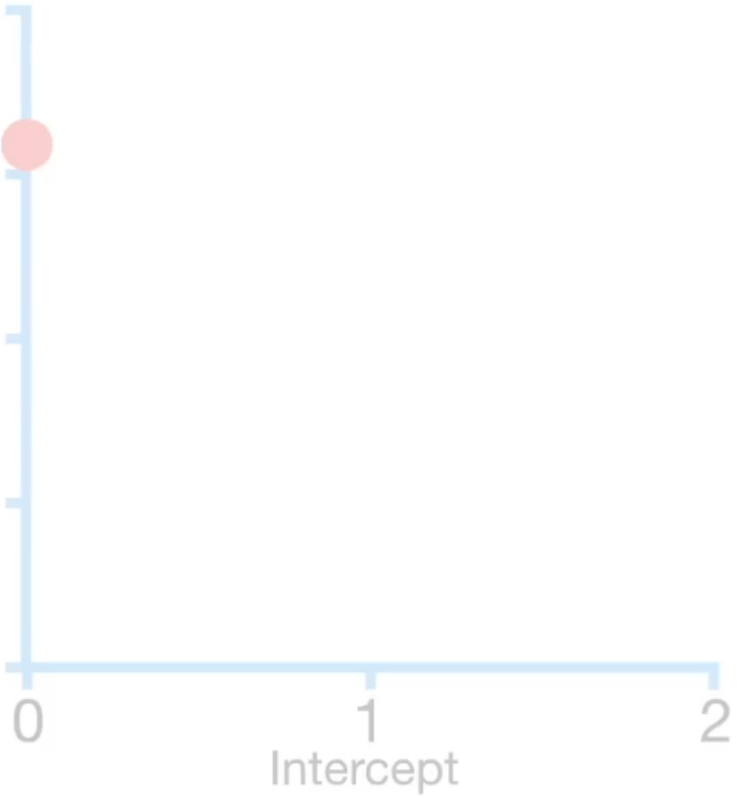
Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2



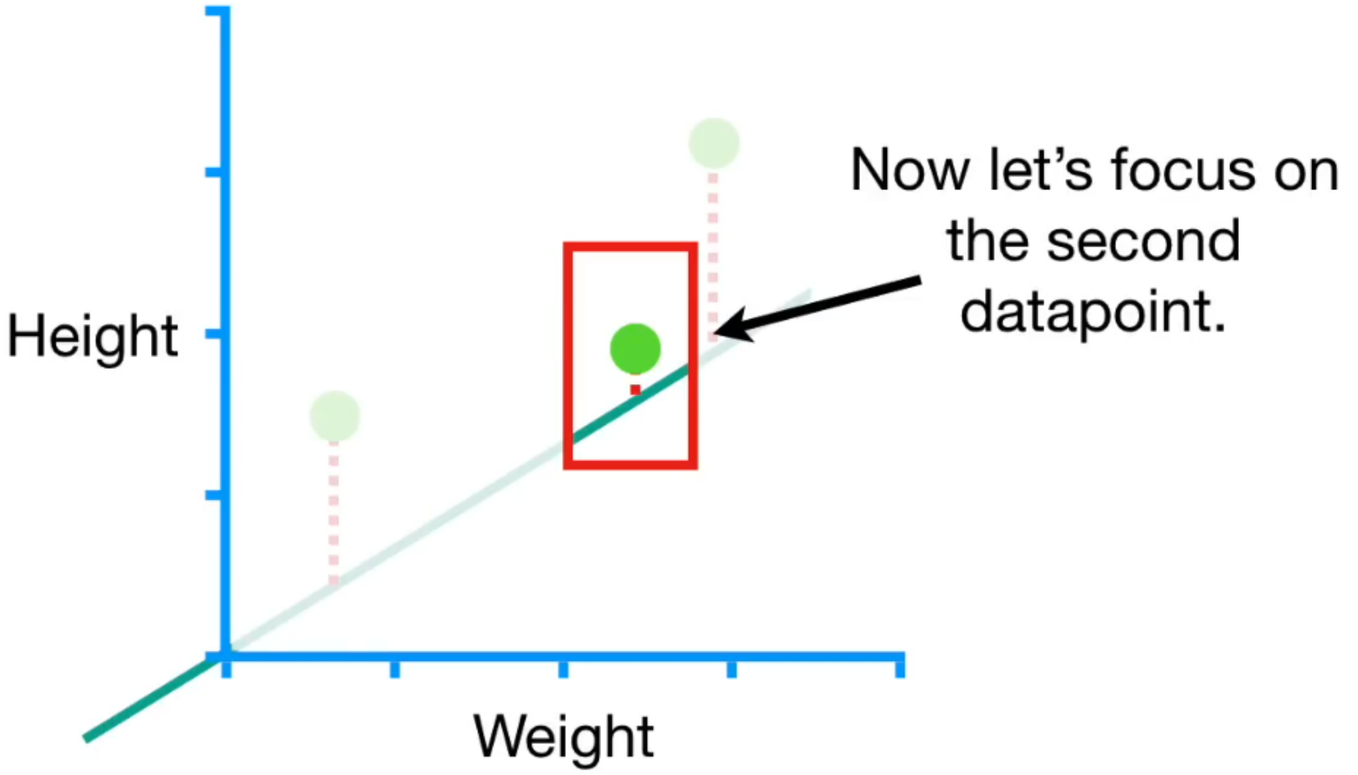
Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

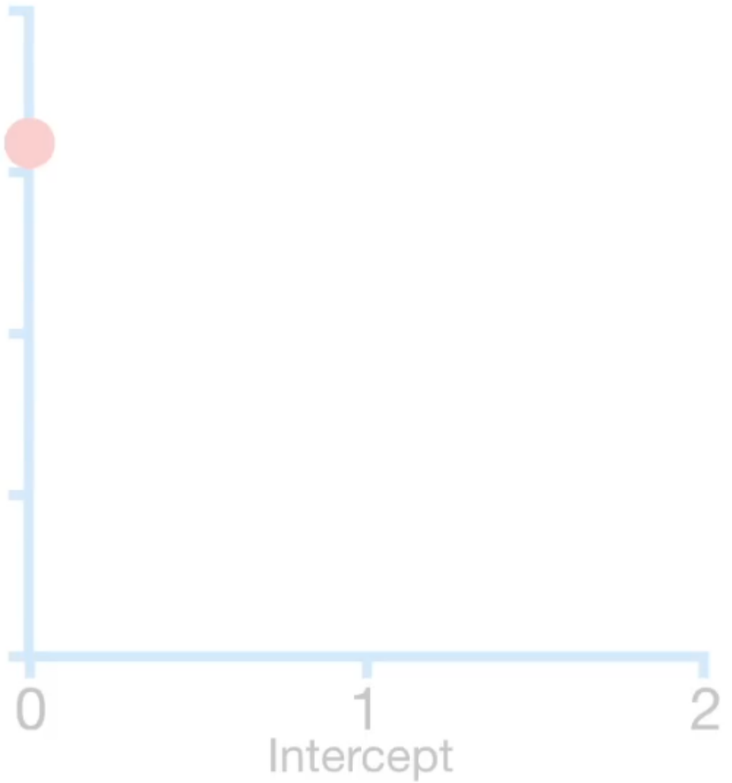


NOTE: We can now plug in any value for the **Intercept** and get a new **Predicted Height**.



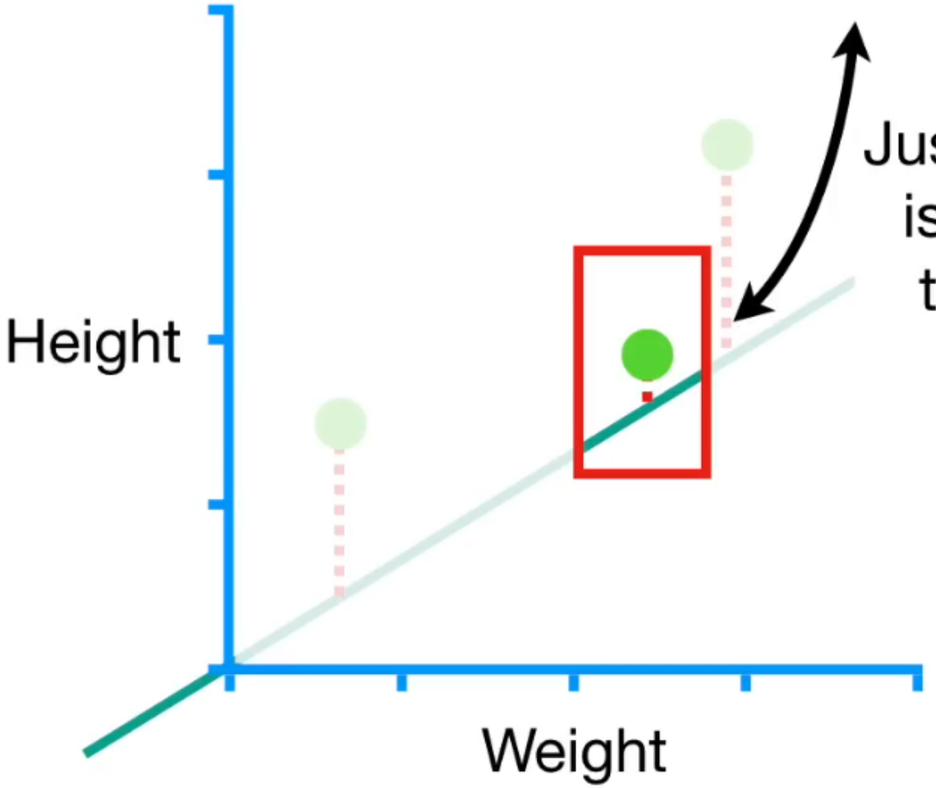
Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2



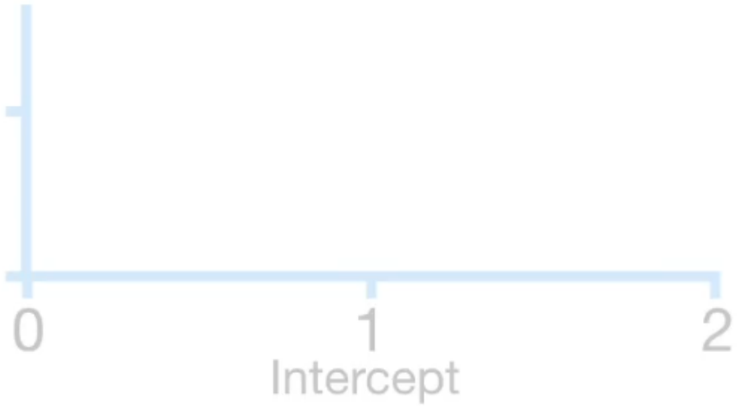


Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

+ (observed - predicted)^2

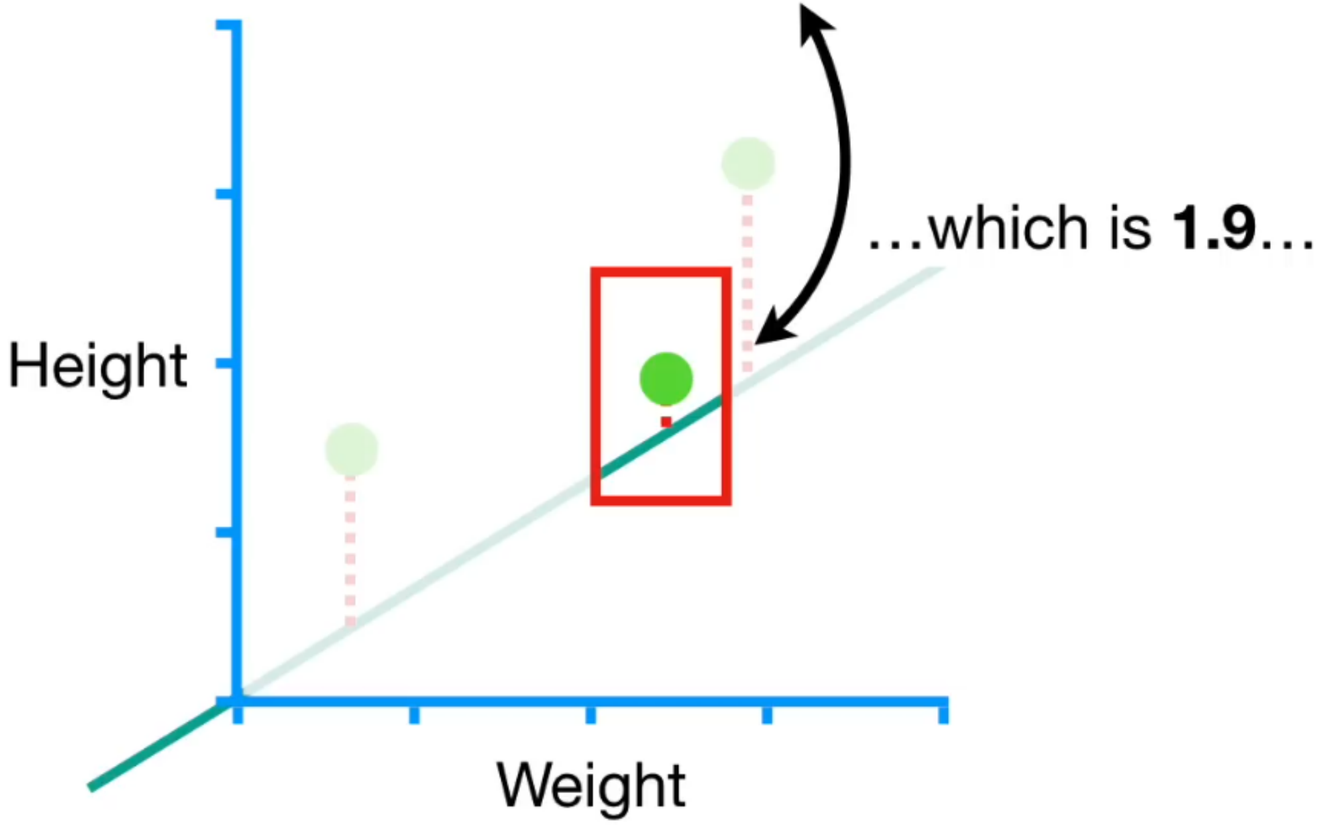


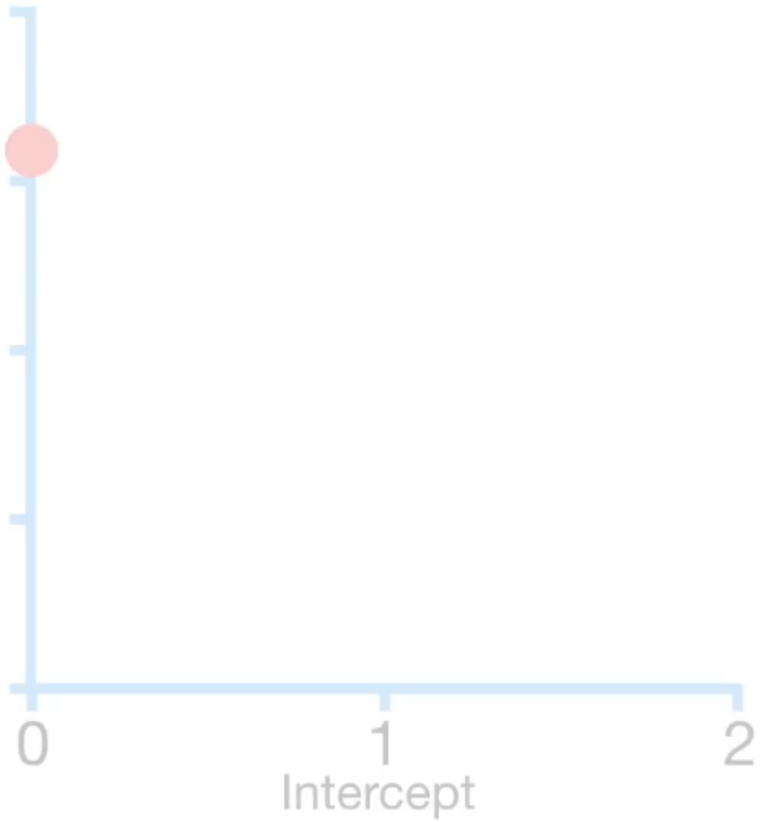
Just like before, the residual is the difference between the **Observed Height**…



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

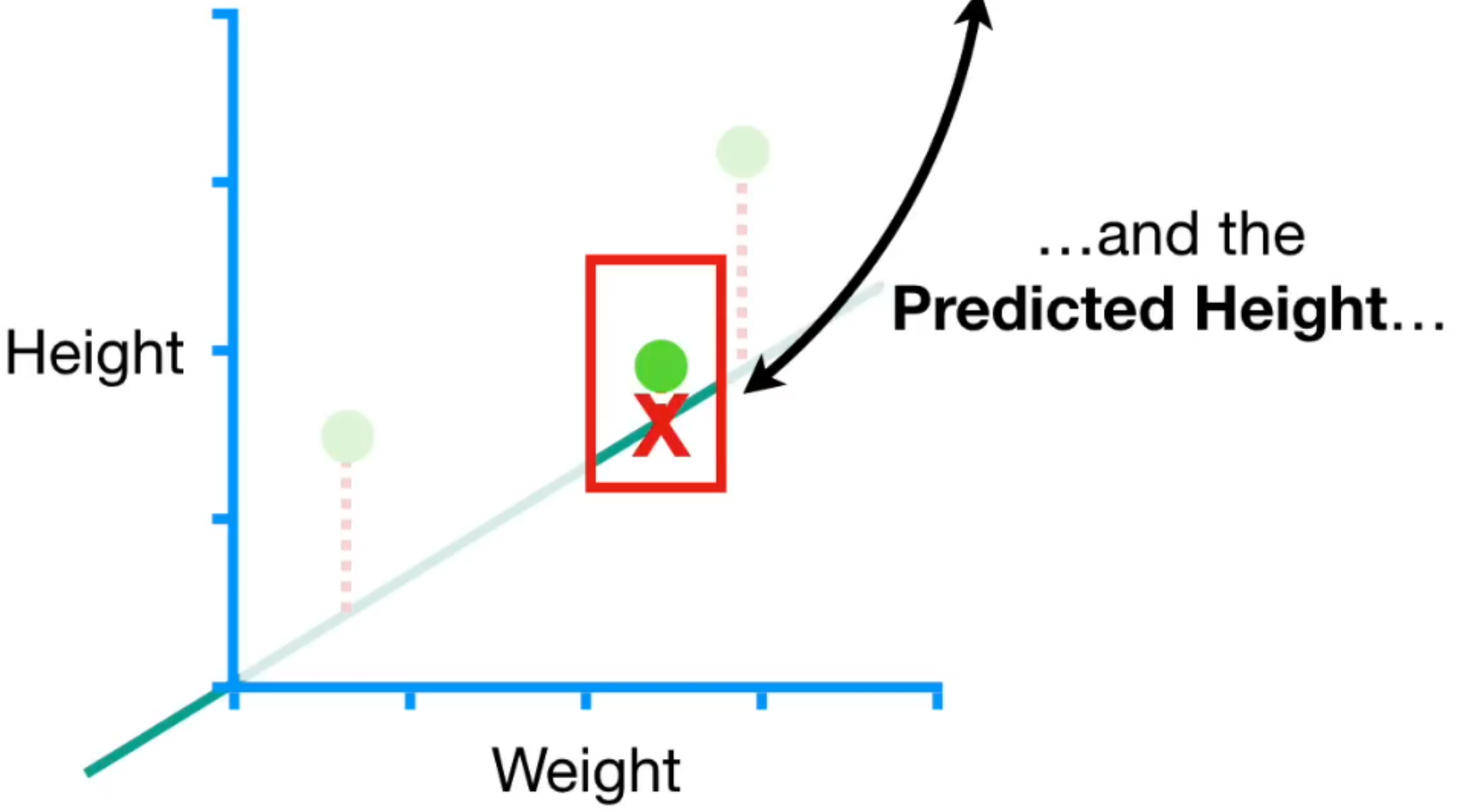
+ (**1.9** - predicted)^2



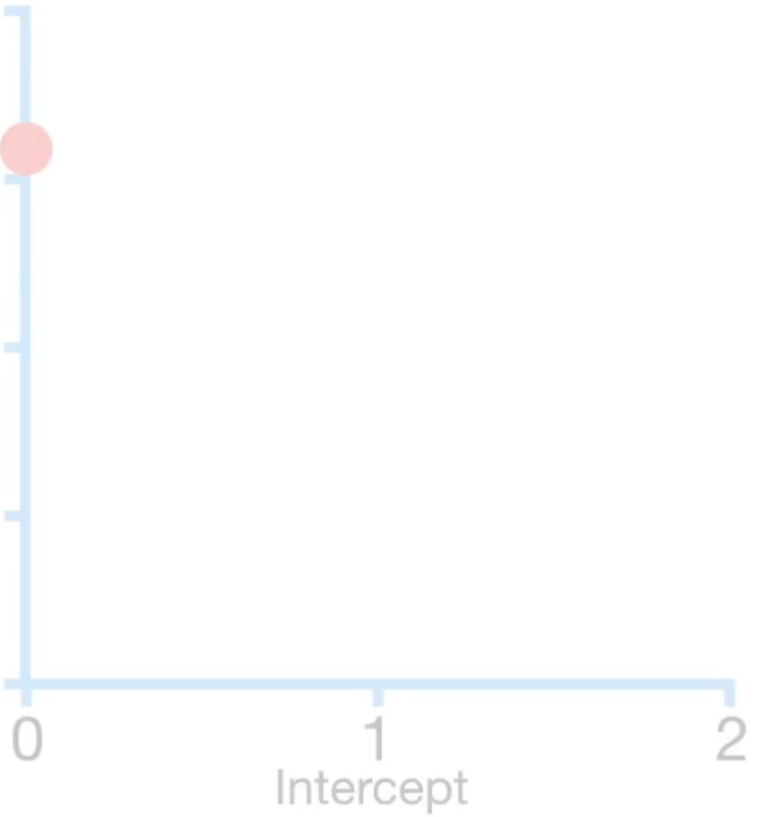


Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

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

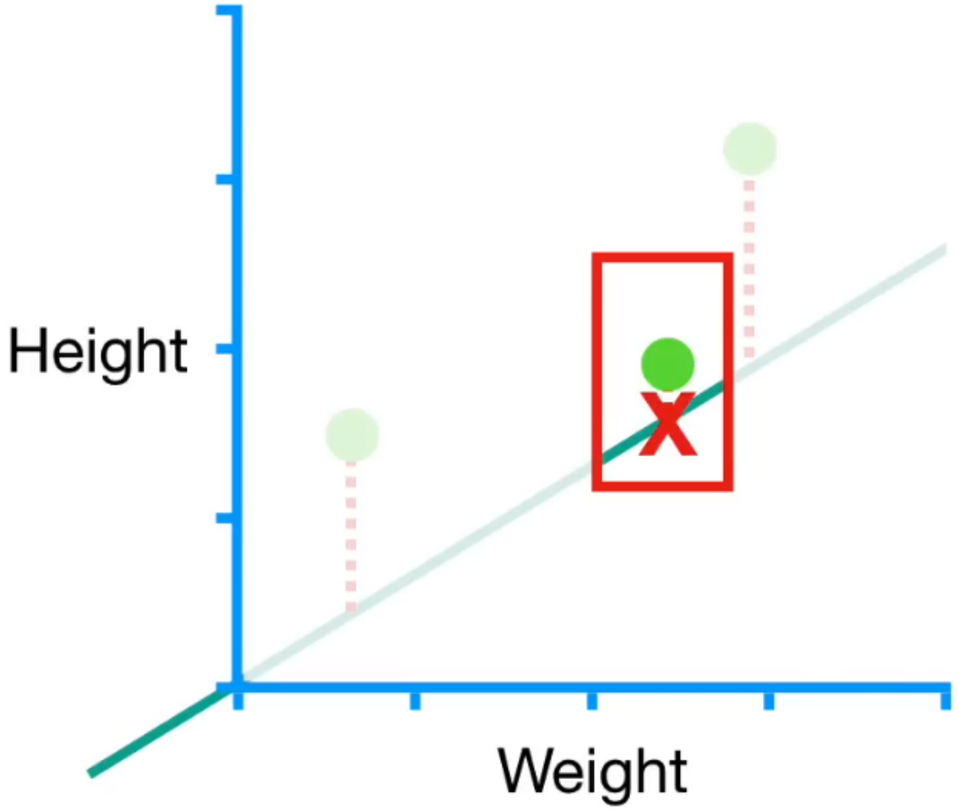


…and the
Predicted Height…

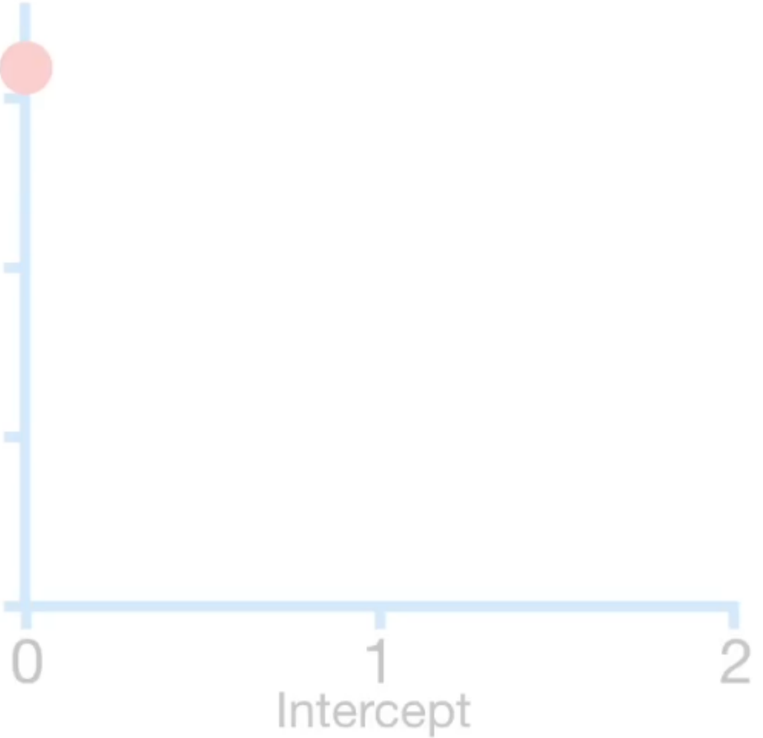


Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

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

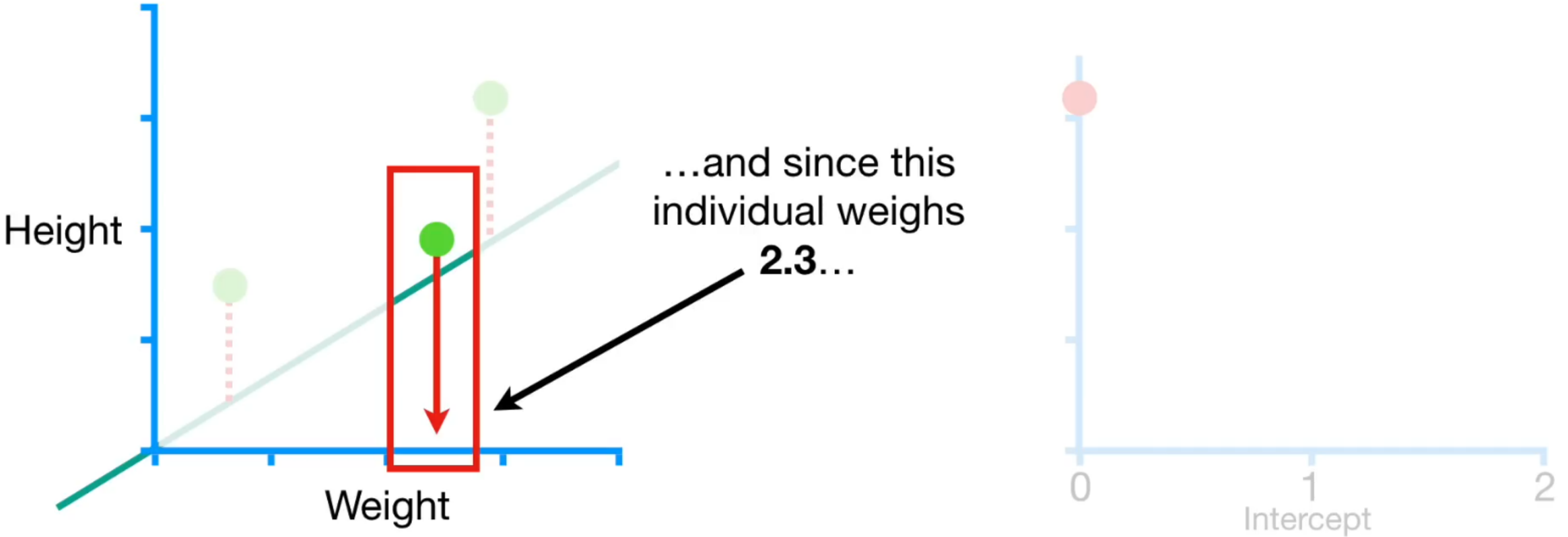


…which comes from the equation for the line…



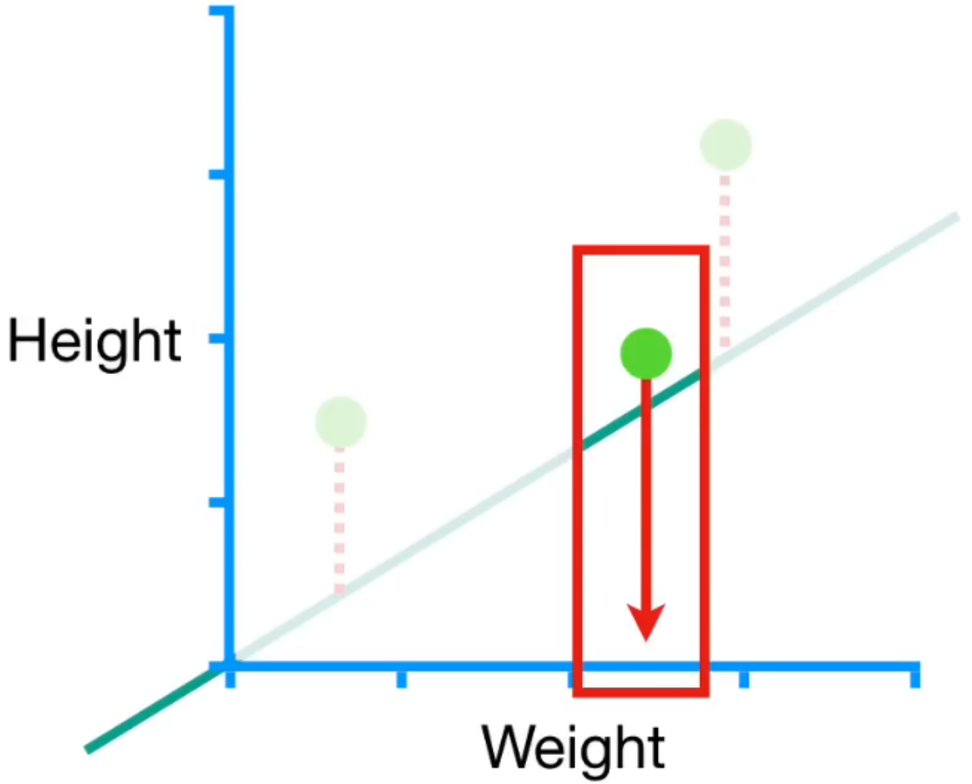
Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

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

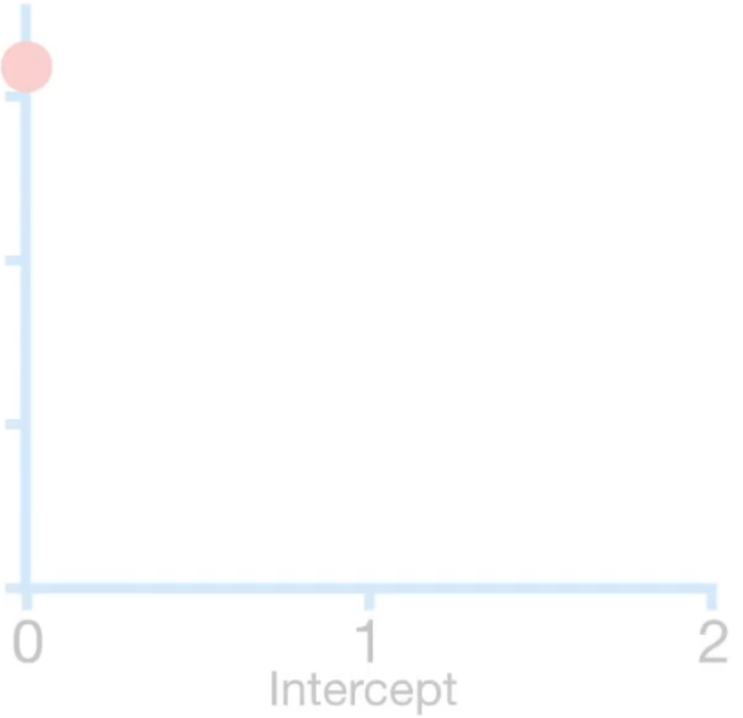


Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

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

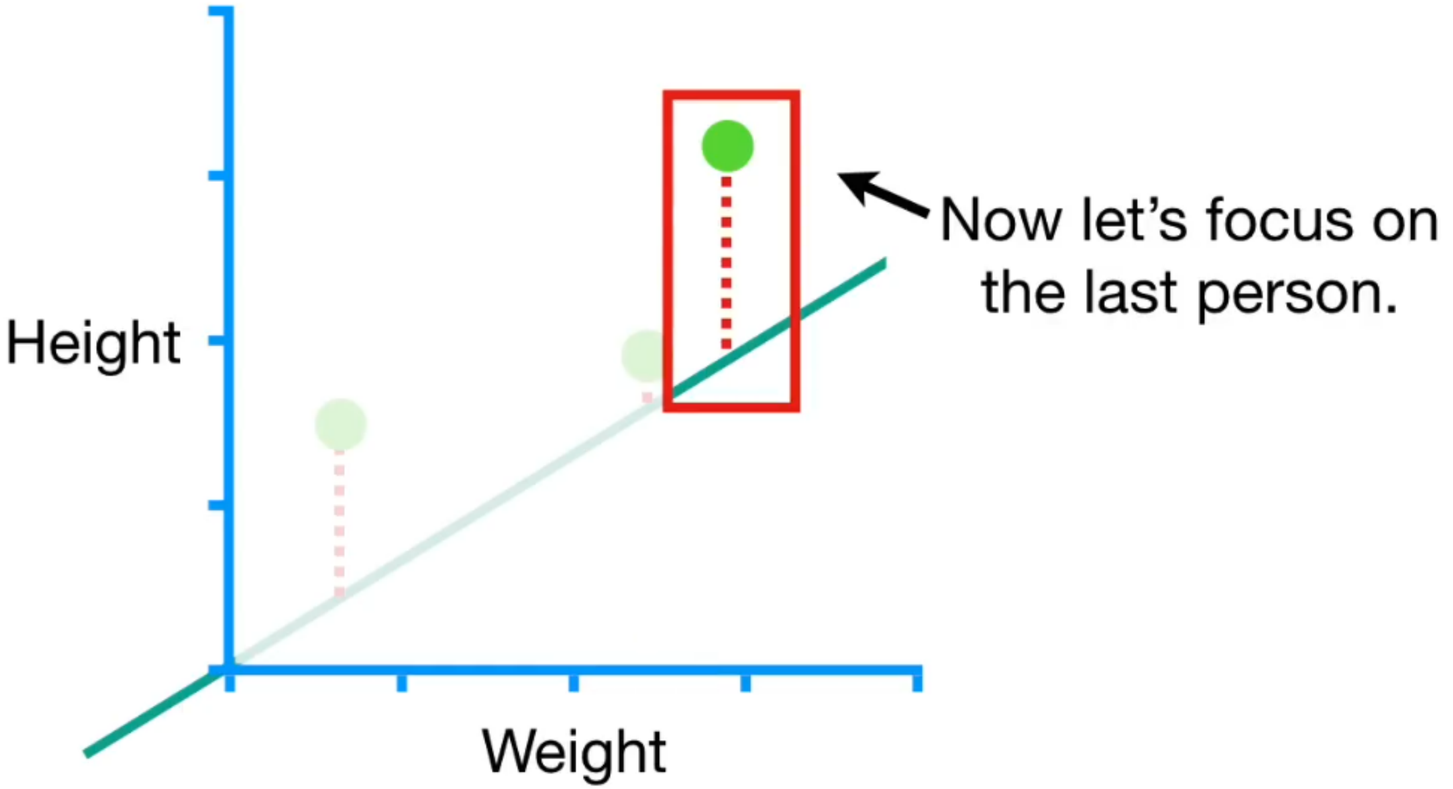


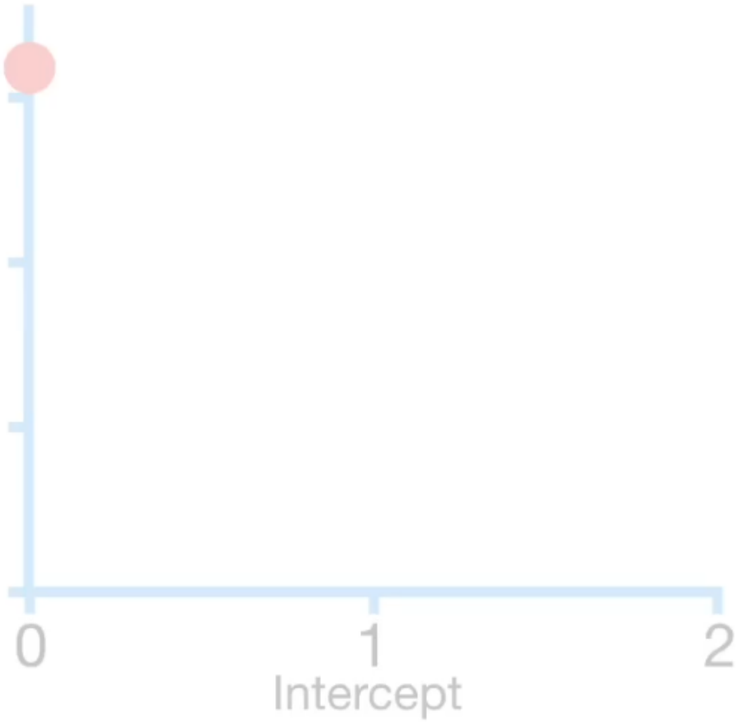
…we replace **weight** with **2.3**.



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

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

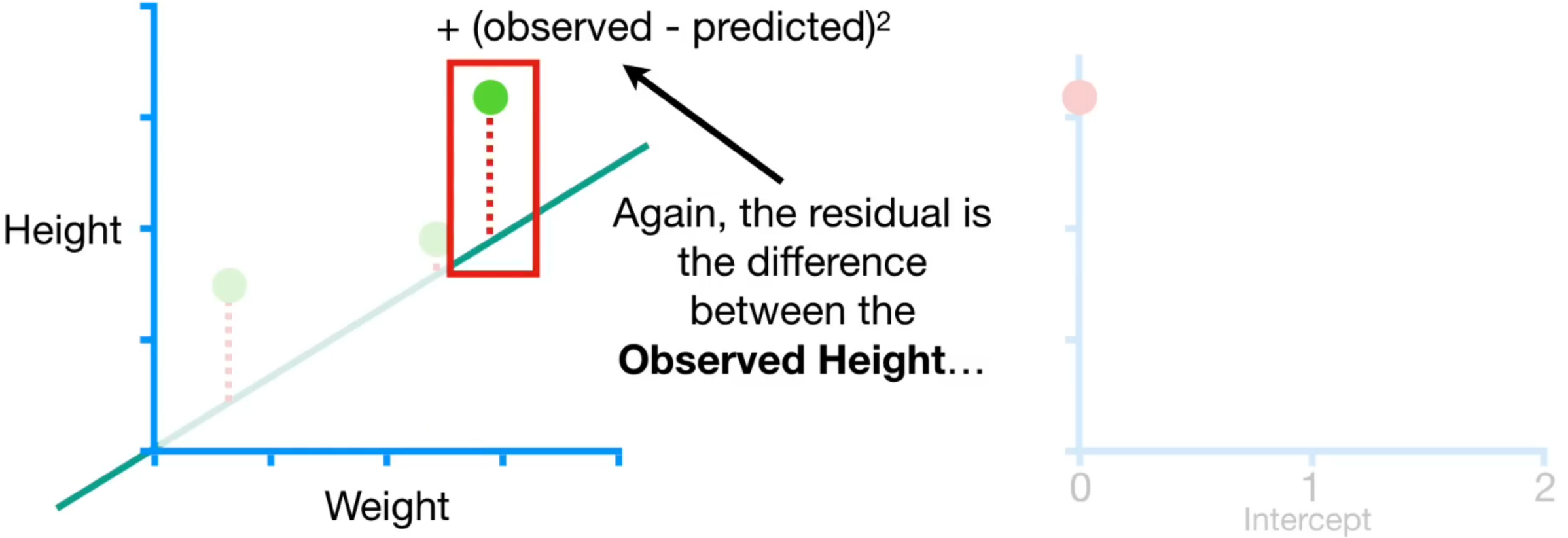




Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

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

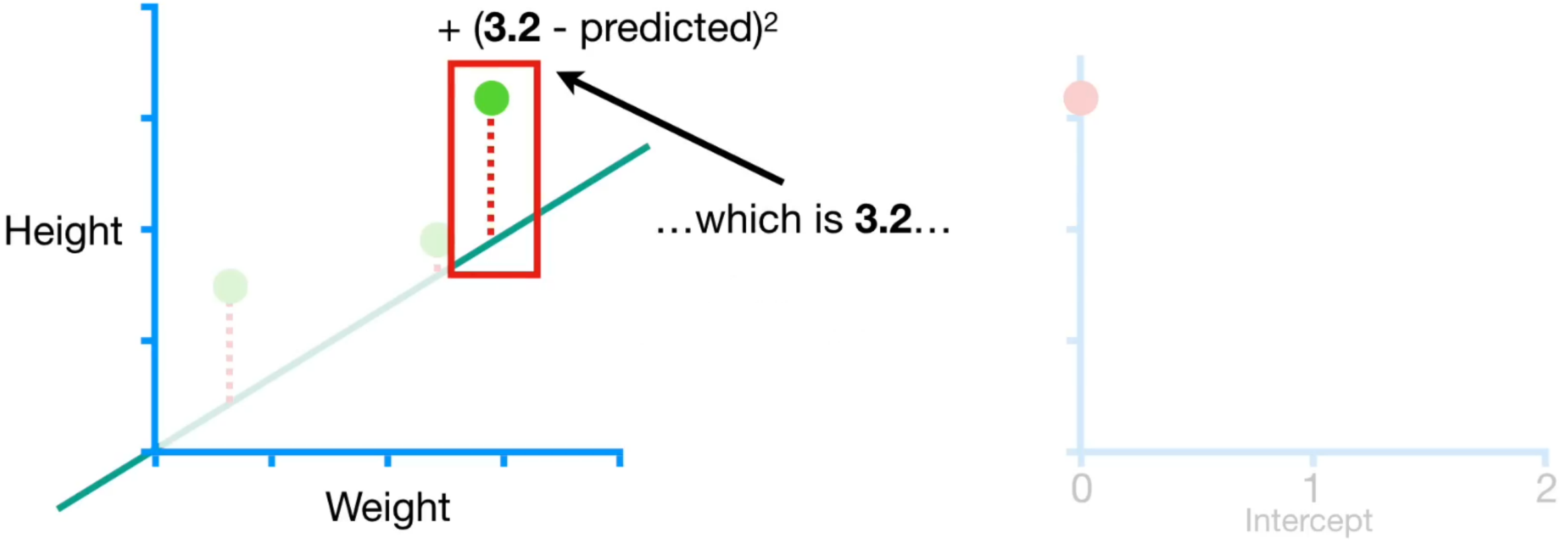
+ (observed - predicted)^2



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

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

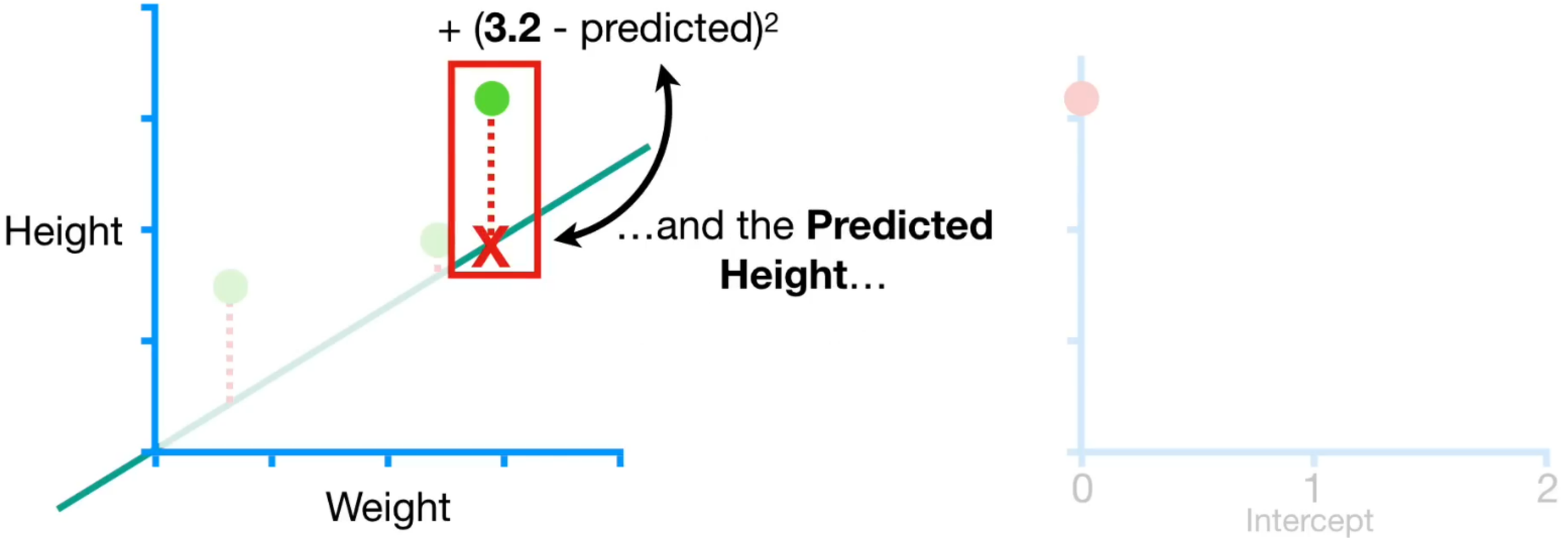
+ (**3.2** - predicted)^2



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

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

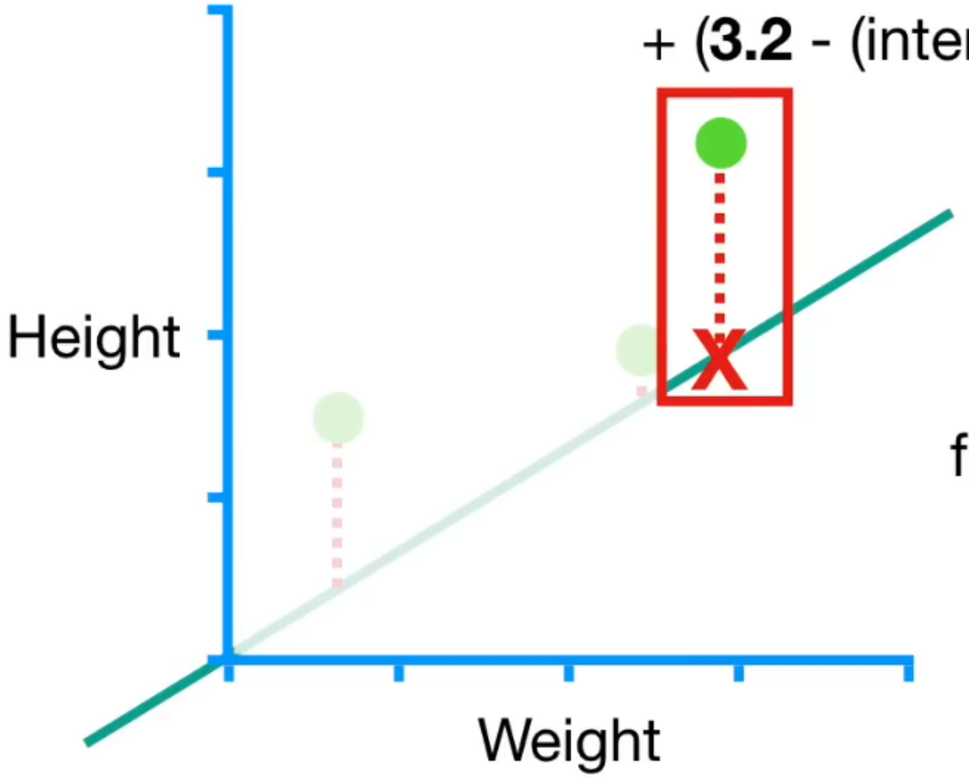
+ (**3.2** - predicted)^2



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 **weight**))^2

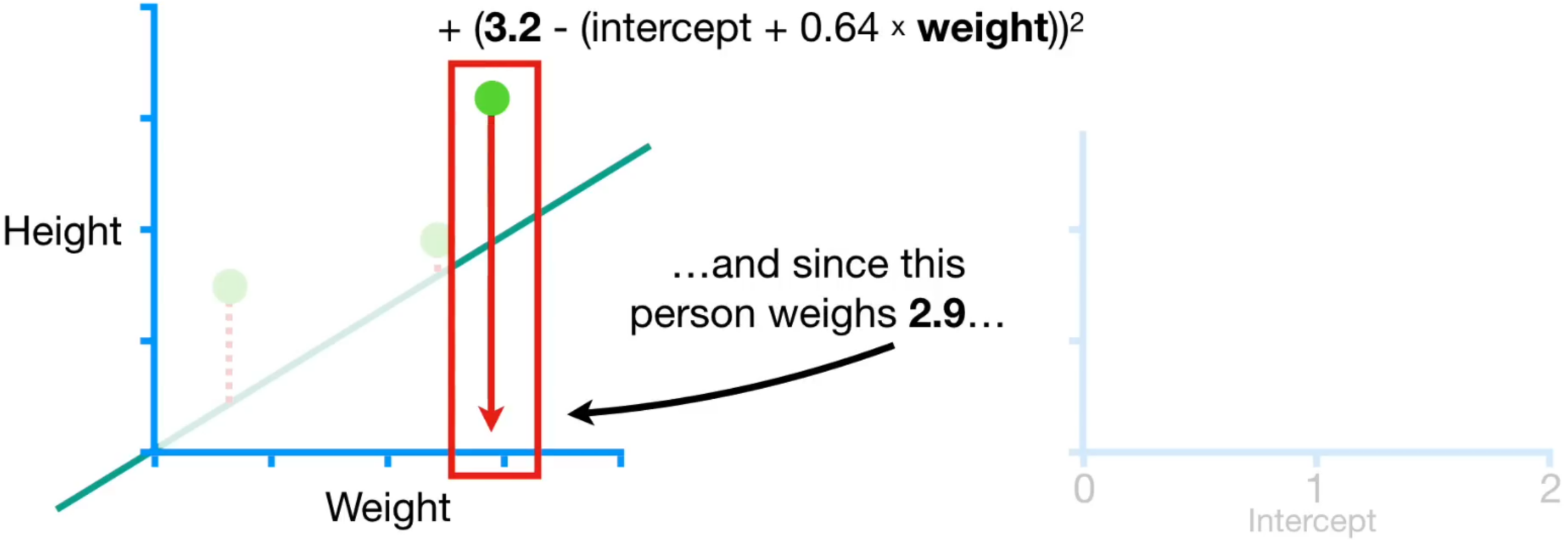


…which comes from the equation for the line…

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

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

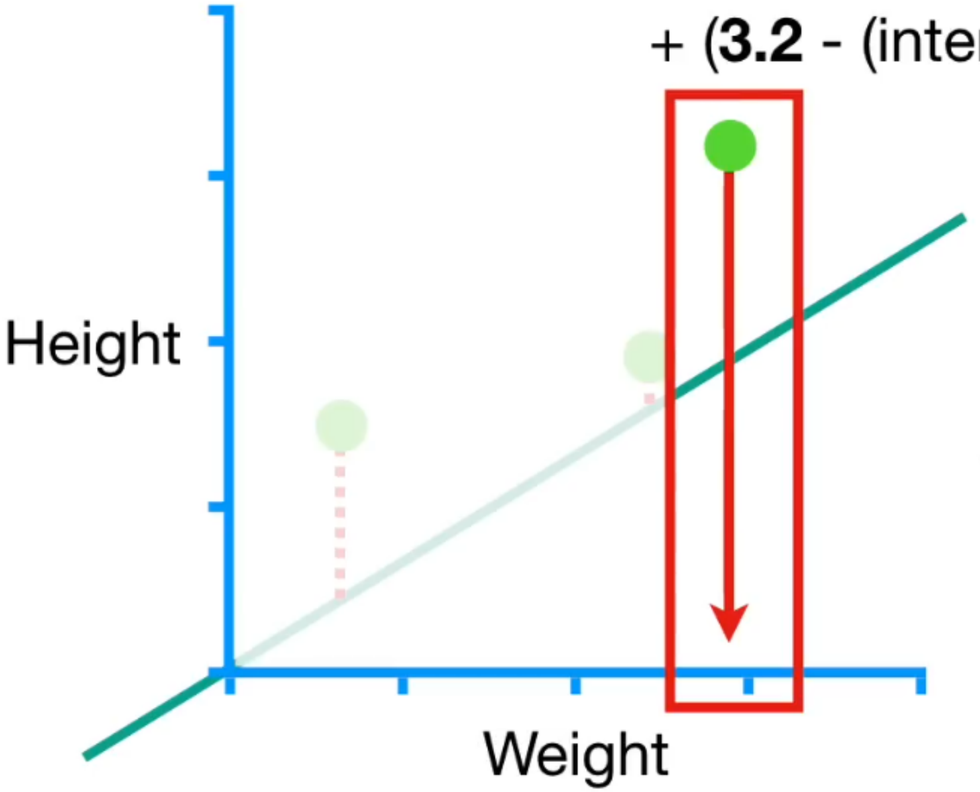
+ (**3.2** - (intercept + 0.64 x **weight**))^2



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

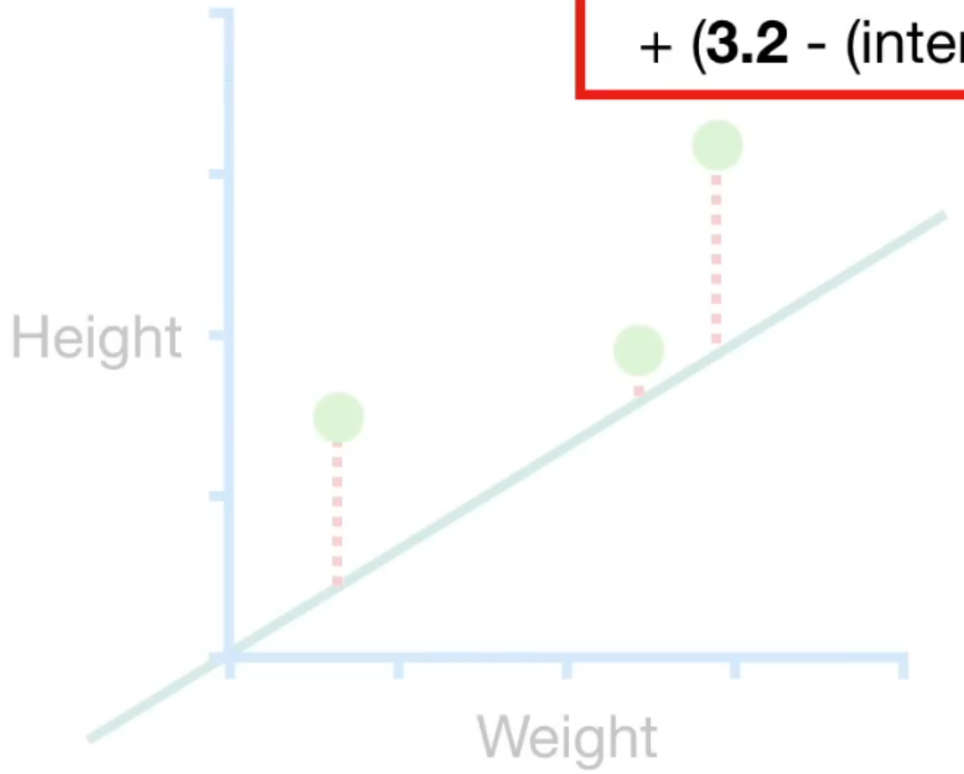


…we'll replace **weight** with **2.9**.

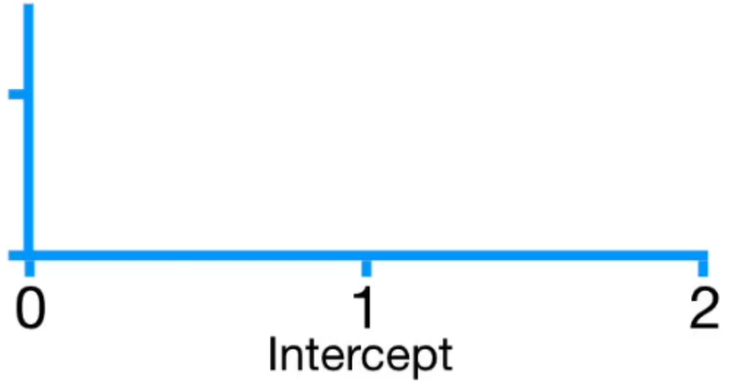
Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

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

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



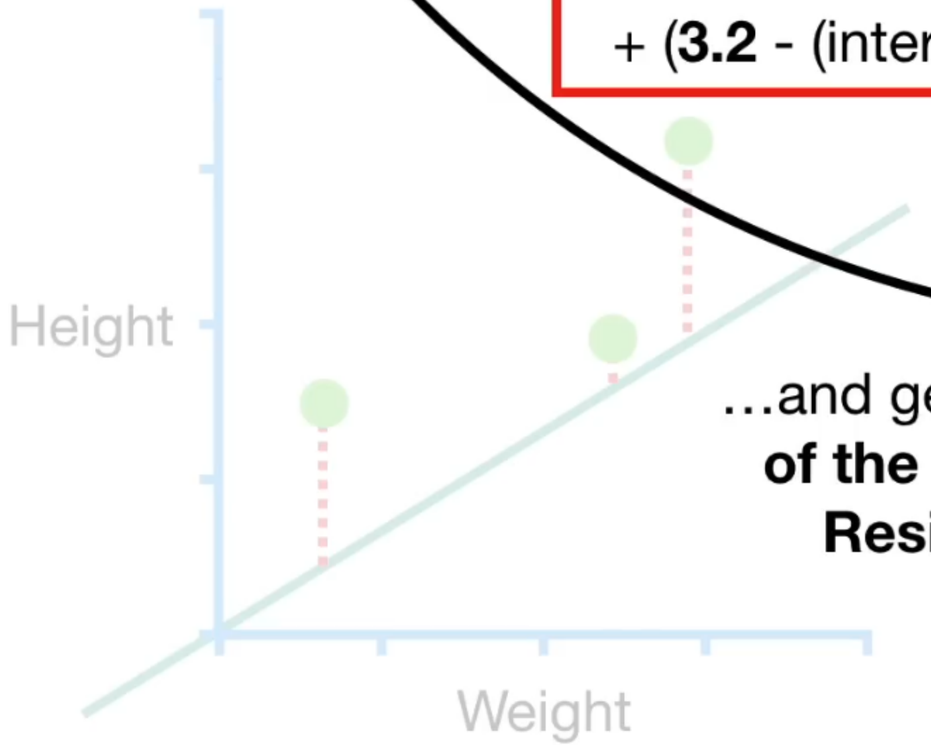
Now we can easily plug in any value for the **intercept**…



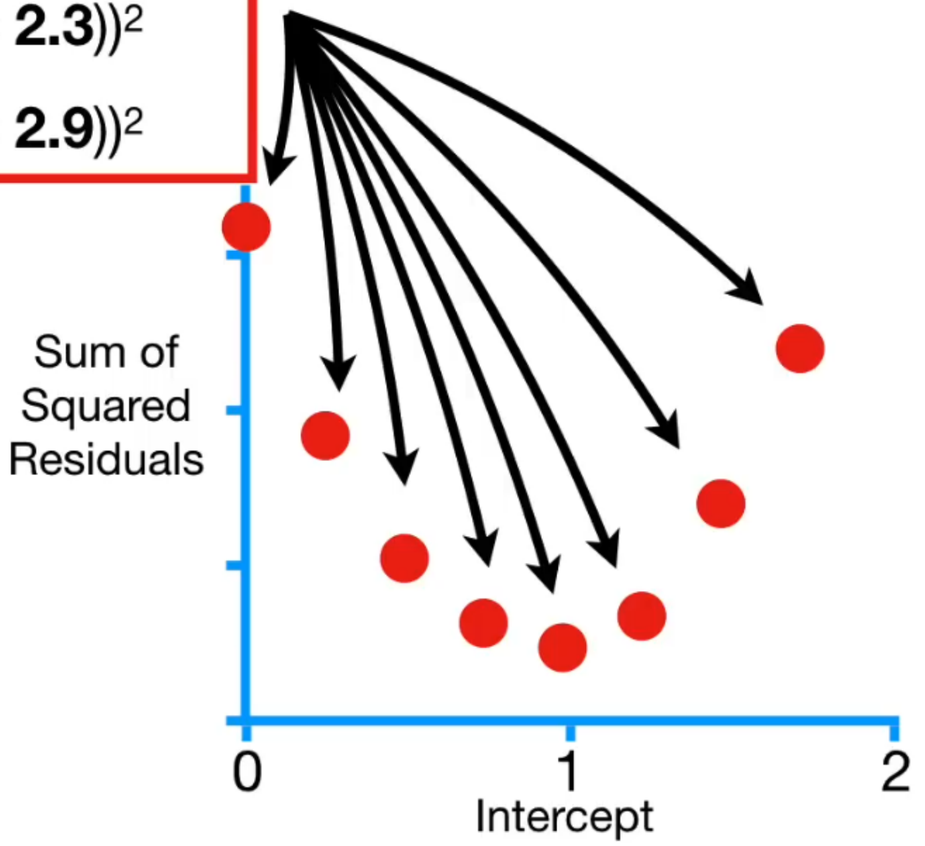
Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

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

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



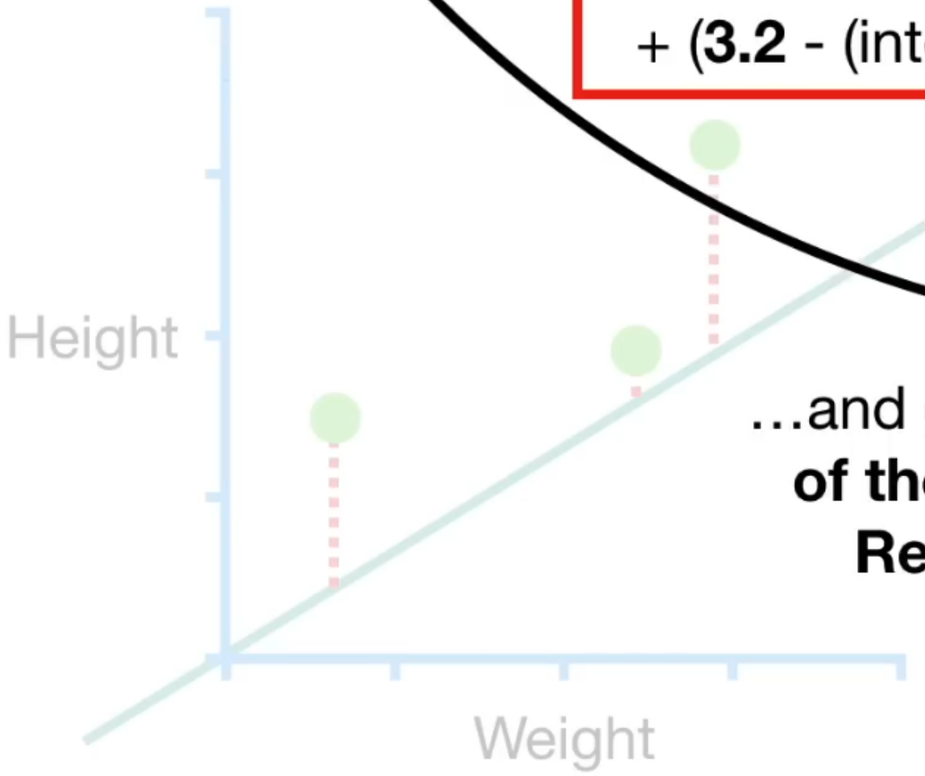
…and get the **Sum of the Squared Residuals**.



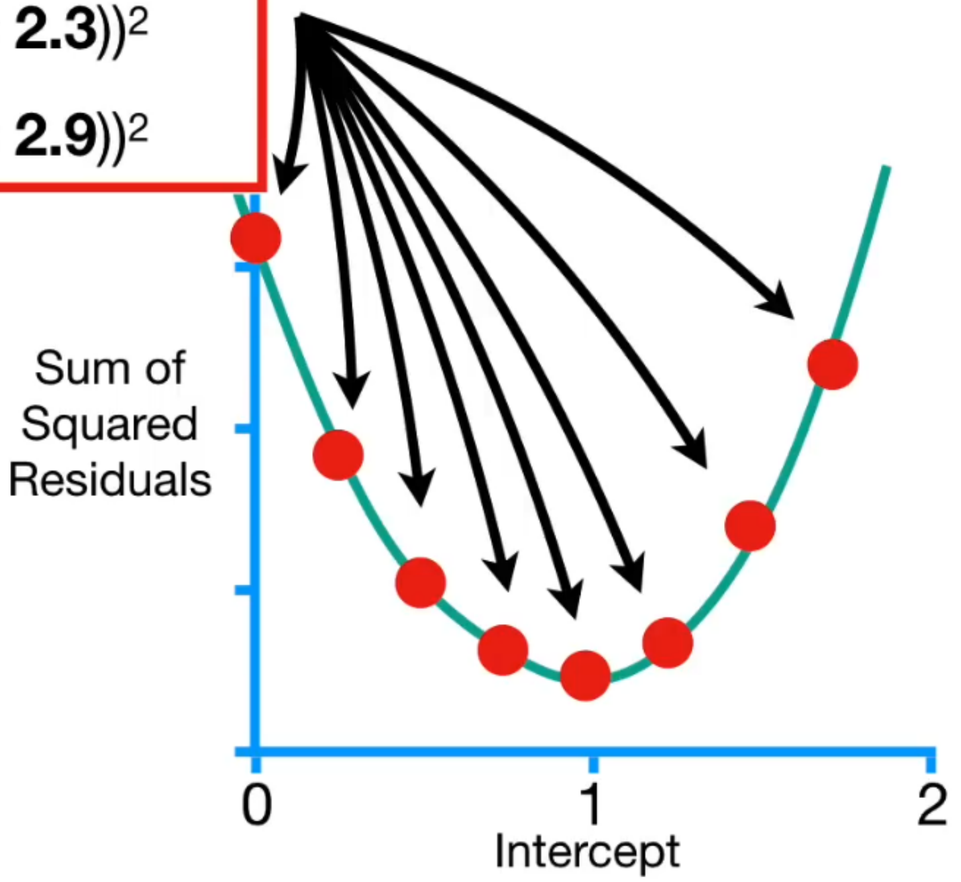
Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

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

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



…and get the **Sum of the Squared Residuals**.

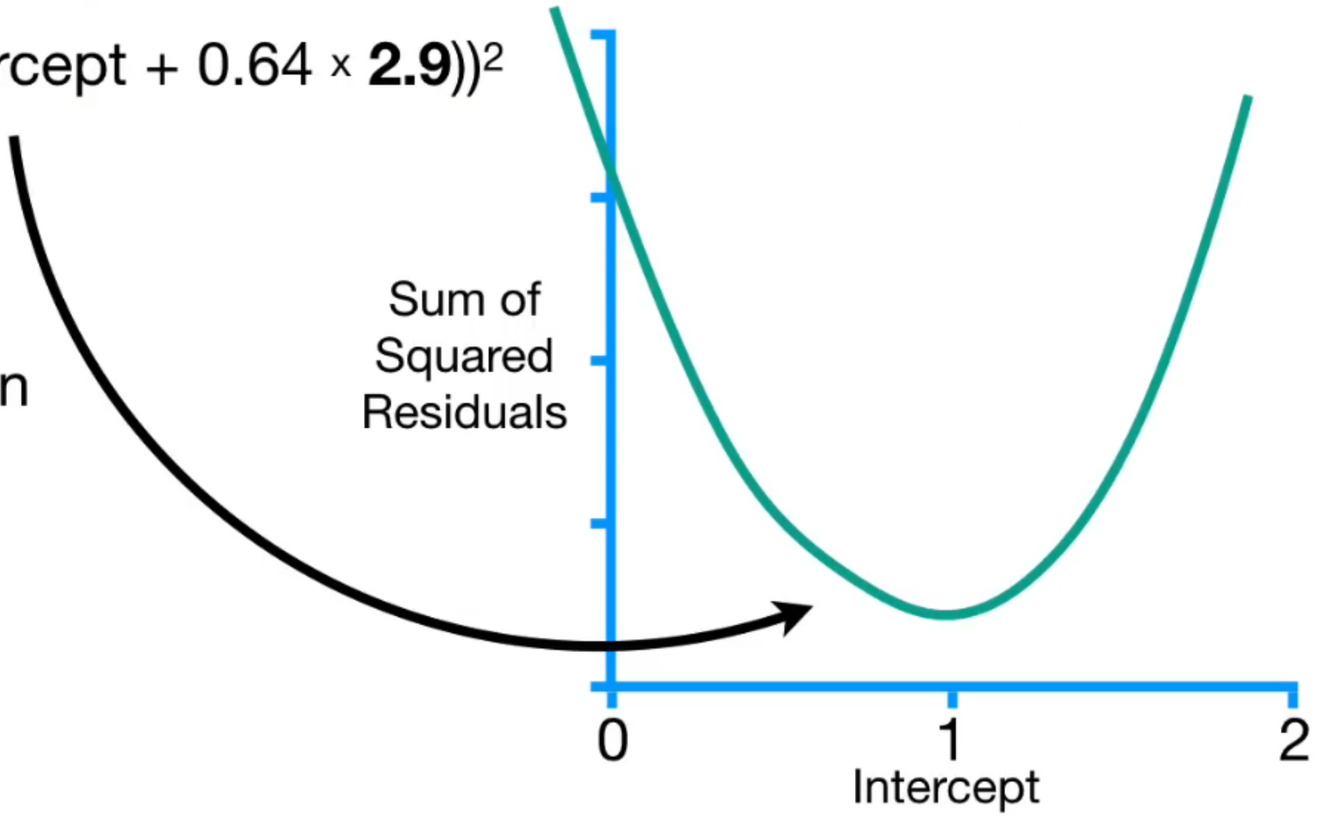


Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

+ (**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…

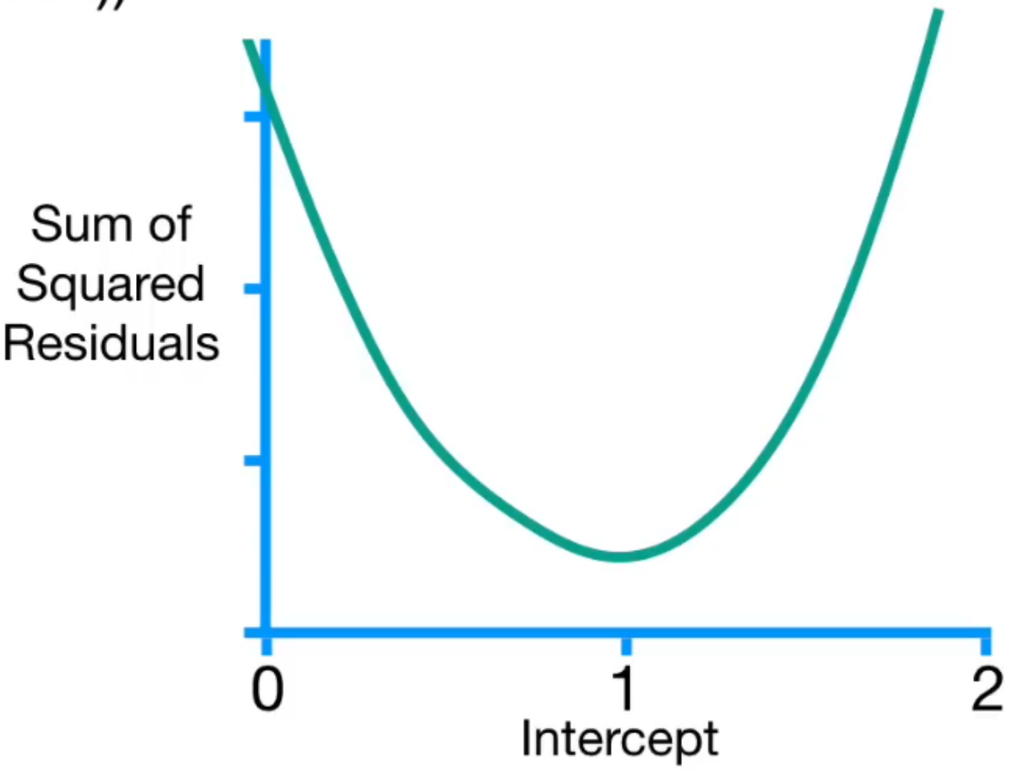


Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

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

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

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



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

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

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

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

Let's start by taking the derivative of the first part.

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

To take the derivative of this, we need to apply

"The Chain Rule"

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

So we start by **moving the square to the front** and **multiply that by the derivative of the stuff inside the parentheses.**

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

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

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

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

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

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

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

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

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

then we simplify by multiplying 2 by -1

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

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

and this is the derivative of the first part
so we will plug it in

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

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

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

Now, we need to take the derivative of the next two parts, and the result will be...

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

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

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

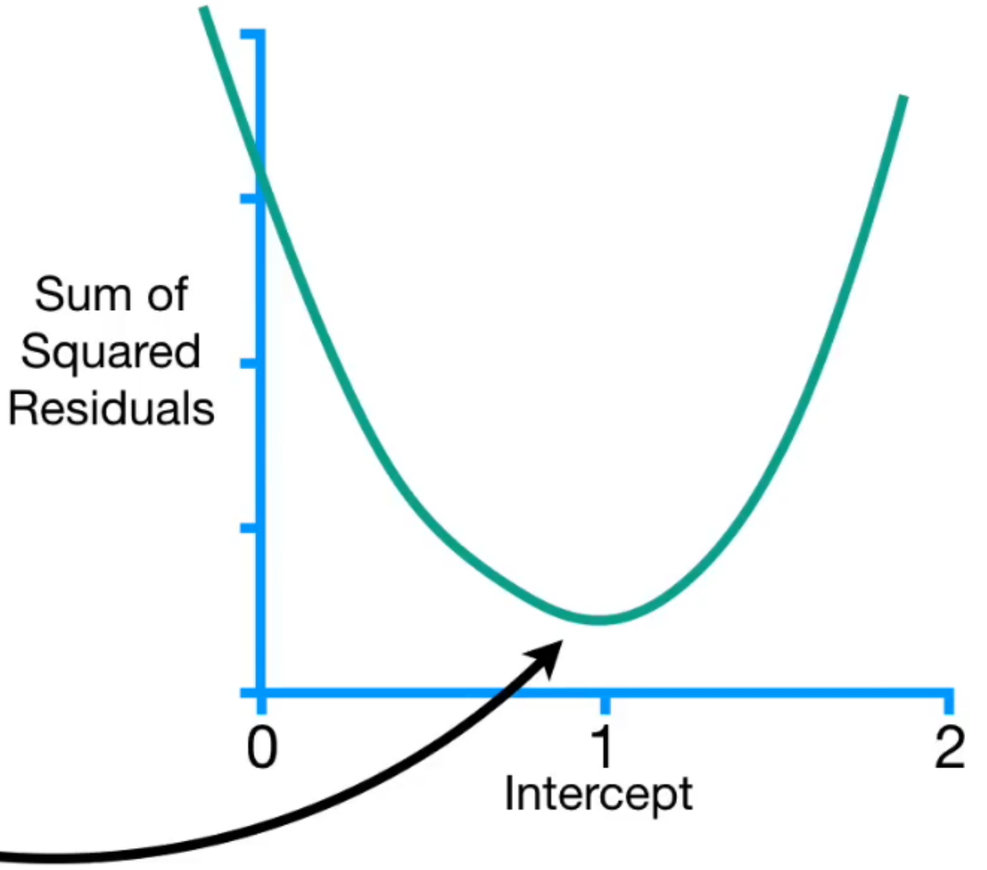
$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

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

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

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

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



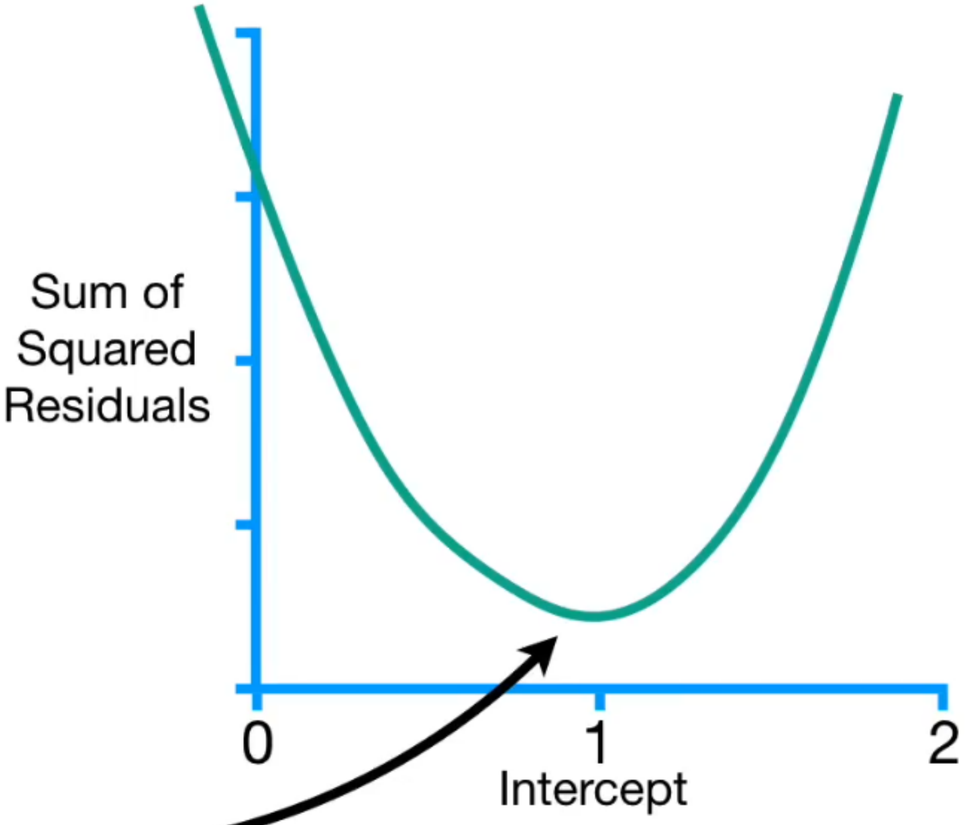
$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

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

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

+ -2(**3.2** - (intercept + 0.64 × **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**.



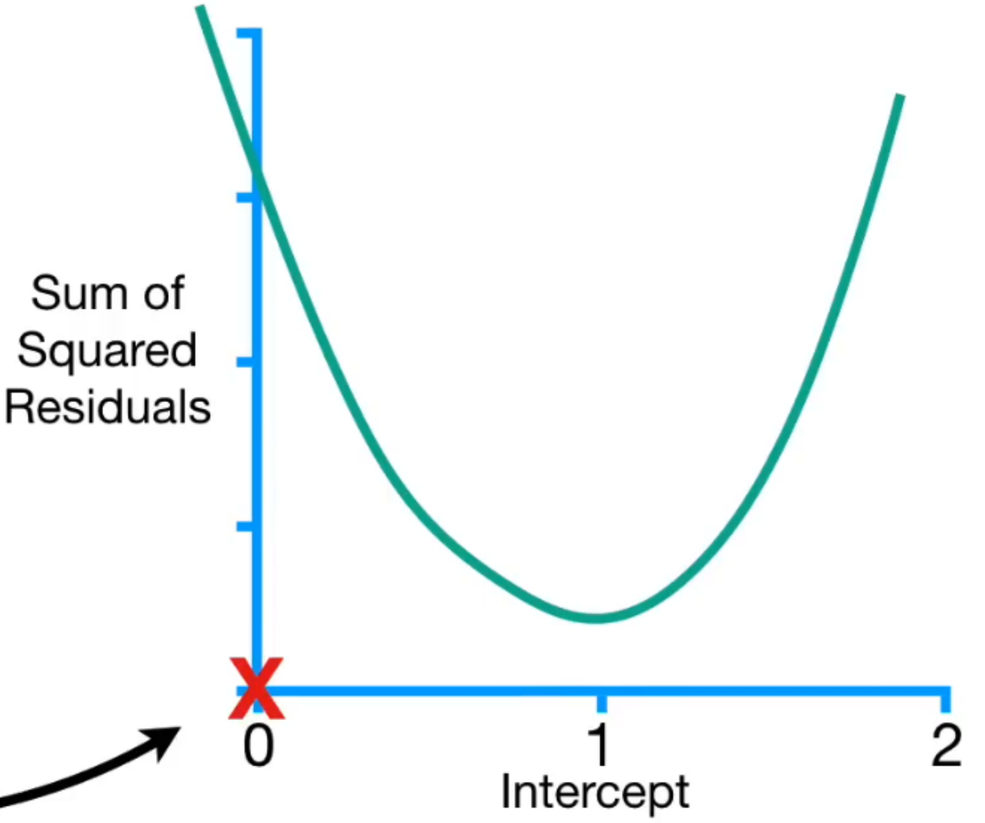
$$\frac{d}{d\,intercept}\text{ Sum of squared residuals} =$$

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

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

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

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

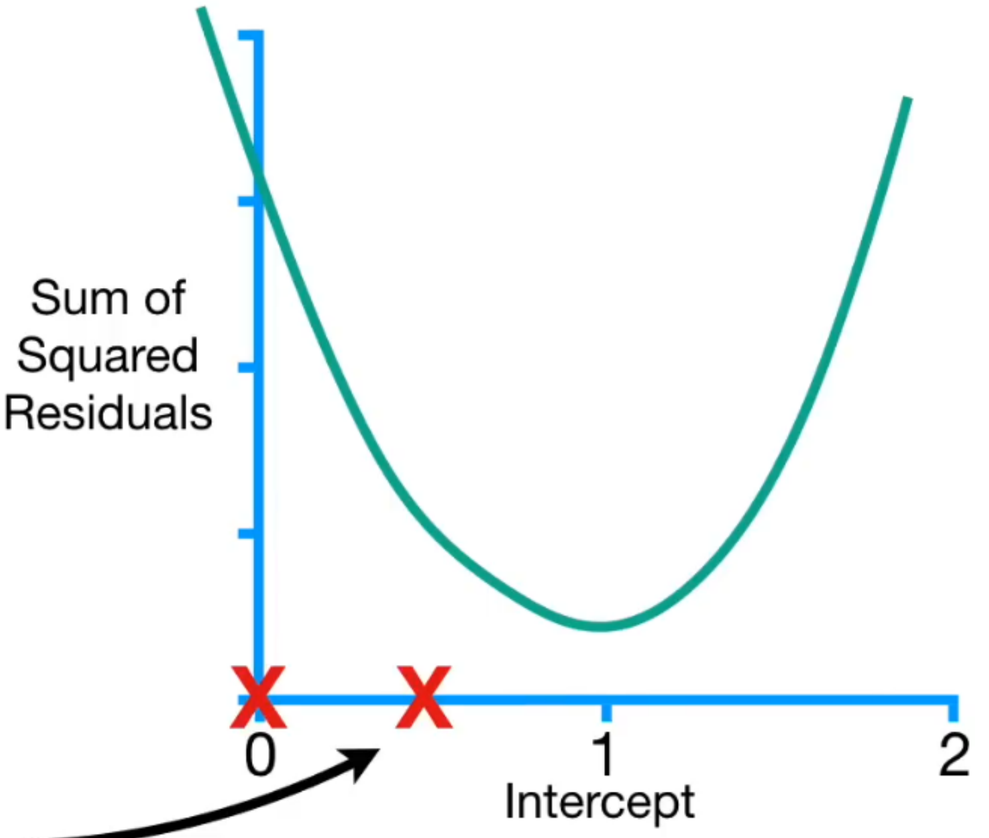


$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

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

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

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



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

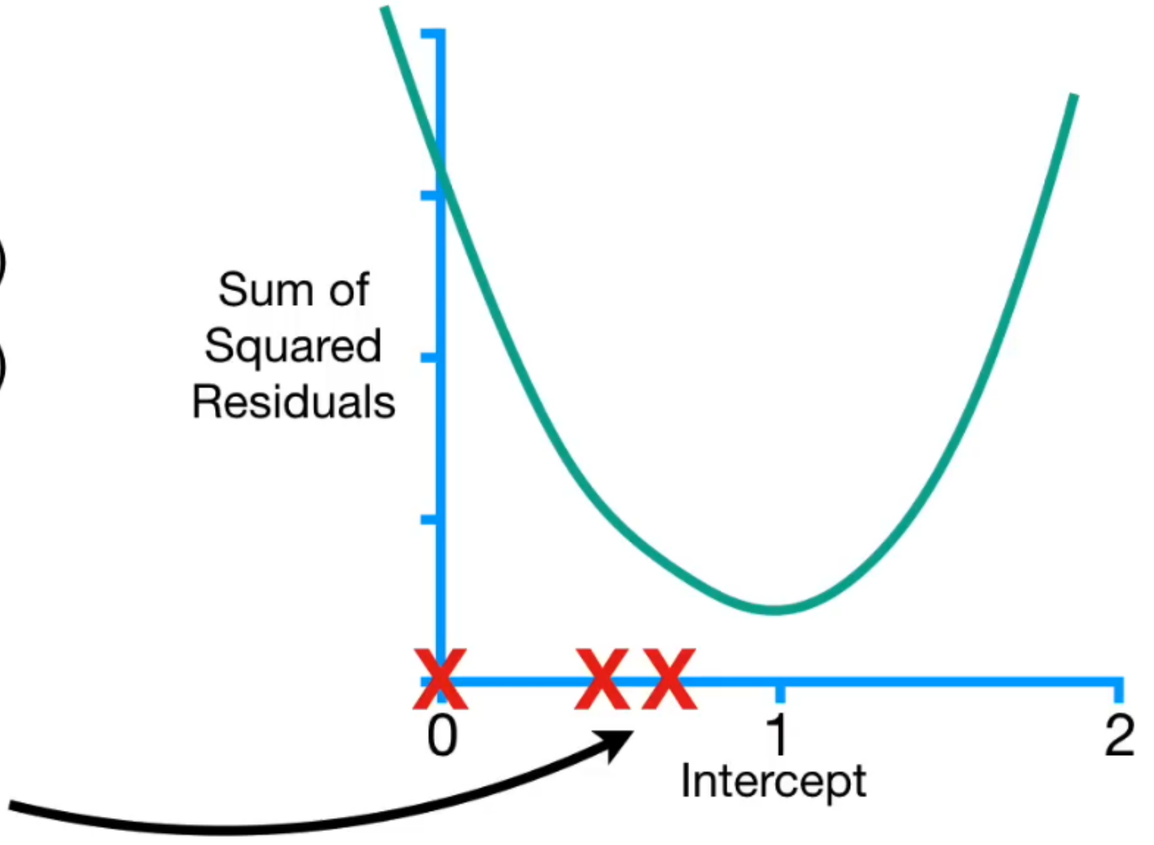
$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

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

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

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

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

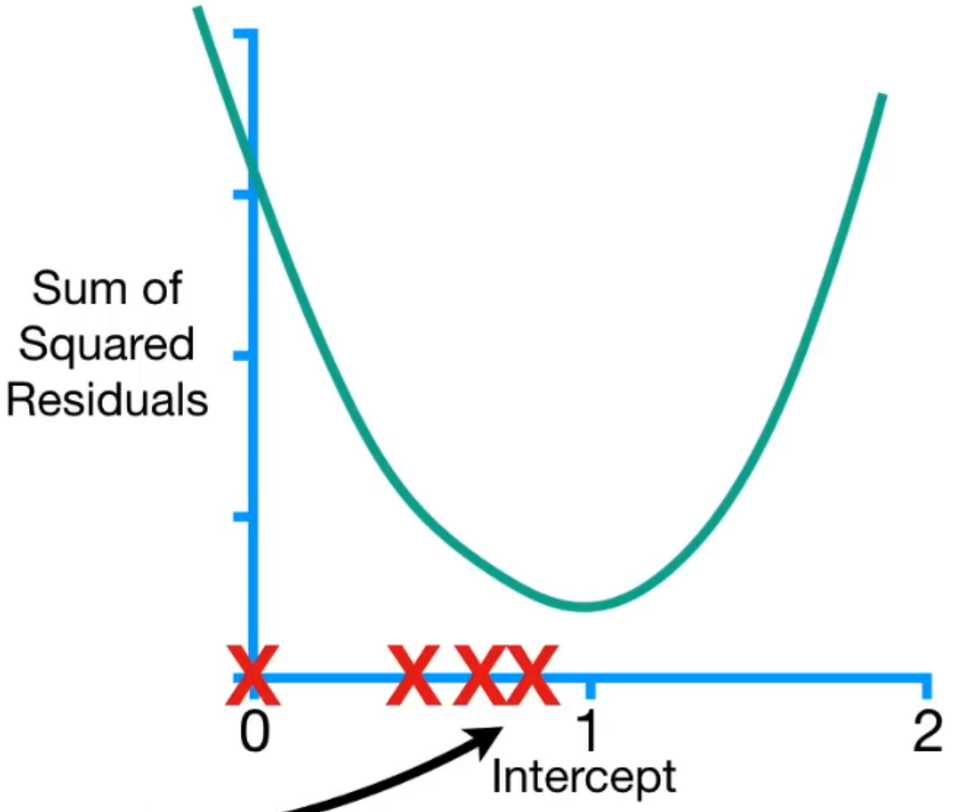


$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

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

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

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



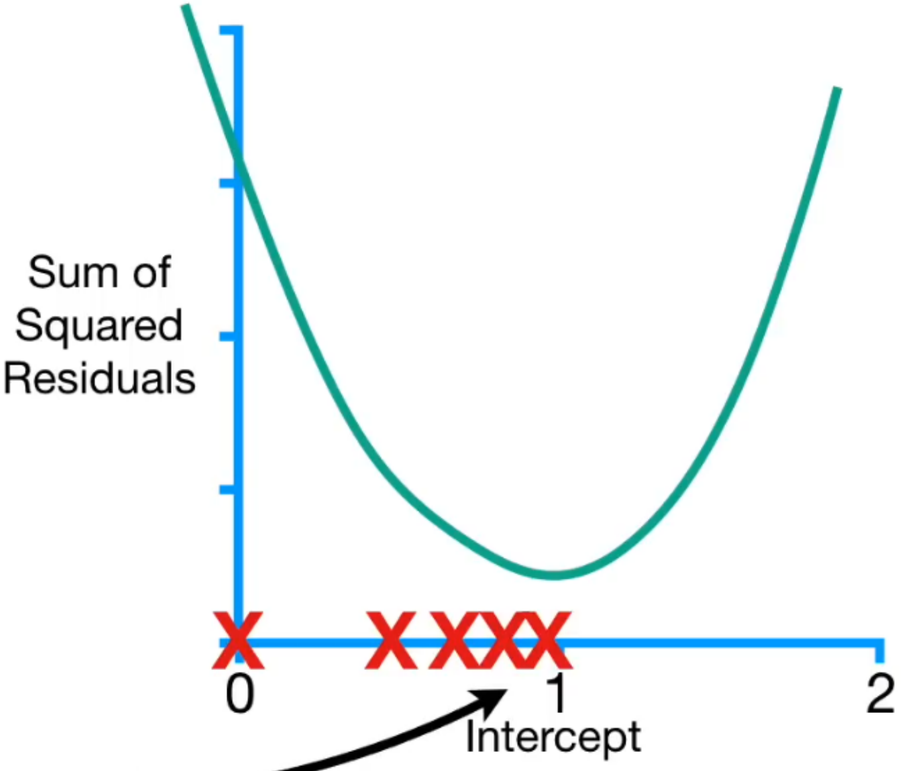
In contrast, **Gradient Descent** finds the minimum value by taking steps from an initial guess until it reaches the best value.

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

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

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

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



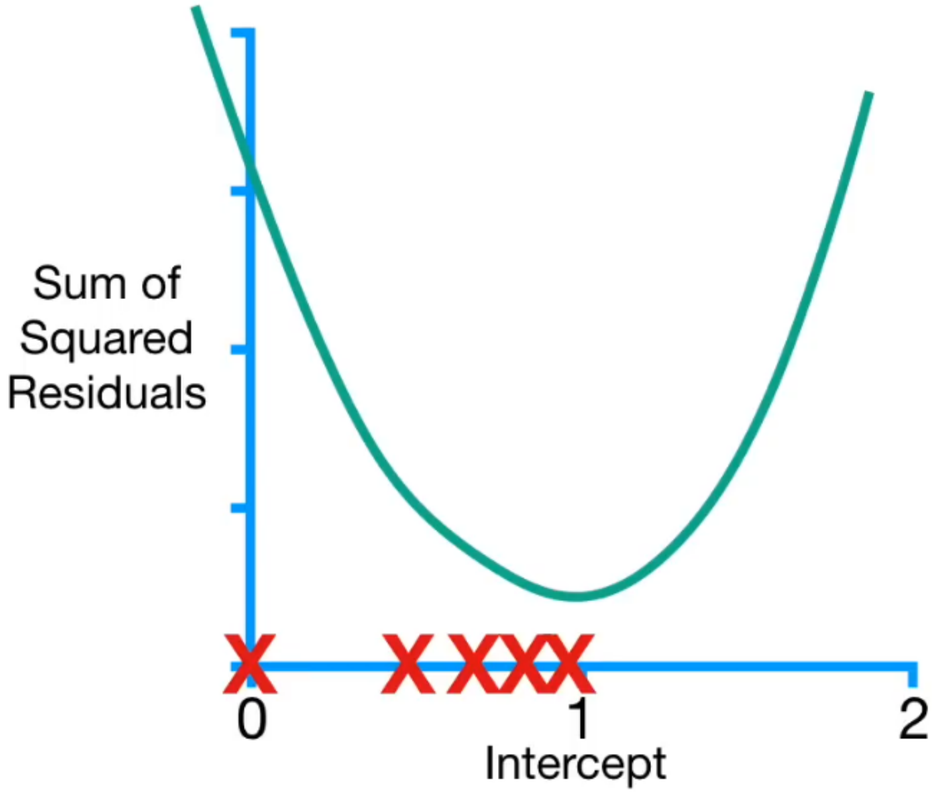
In contrast, **Gradient Descent** finds the minimum value by taking steps from an initial guess until it reaches the best value.

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

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

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

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



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

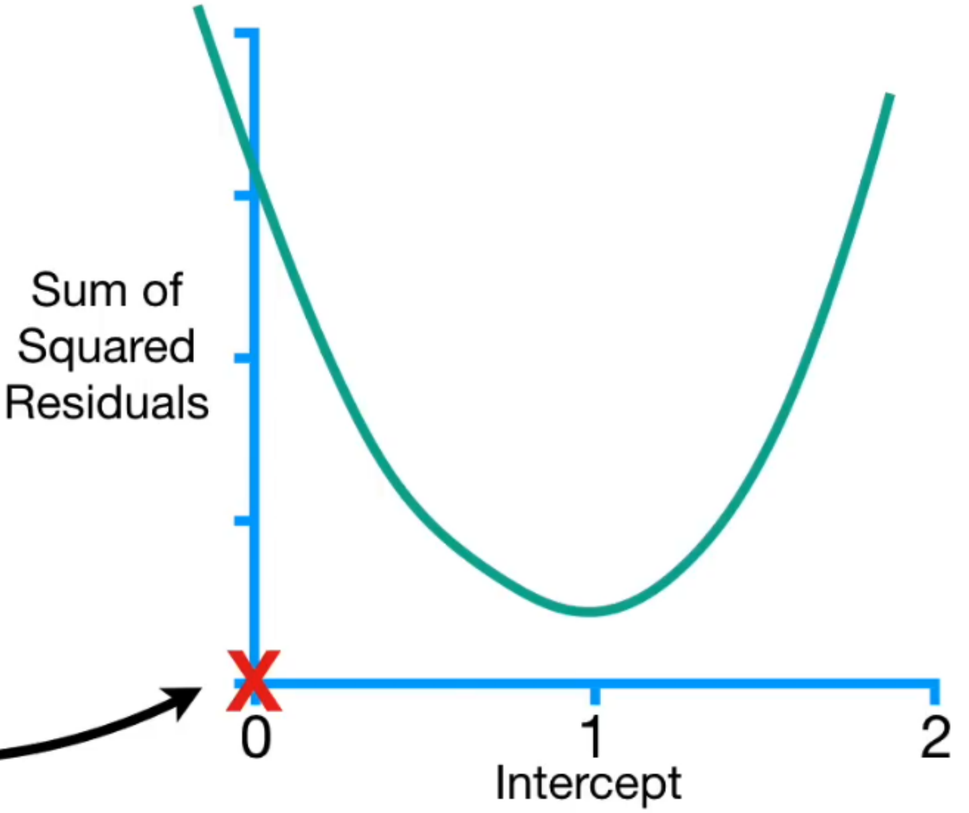
$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

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

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

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

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



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

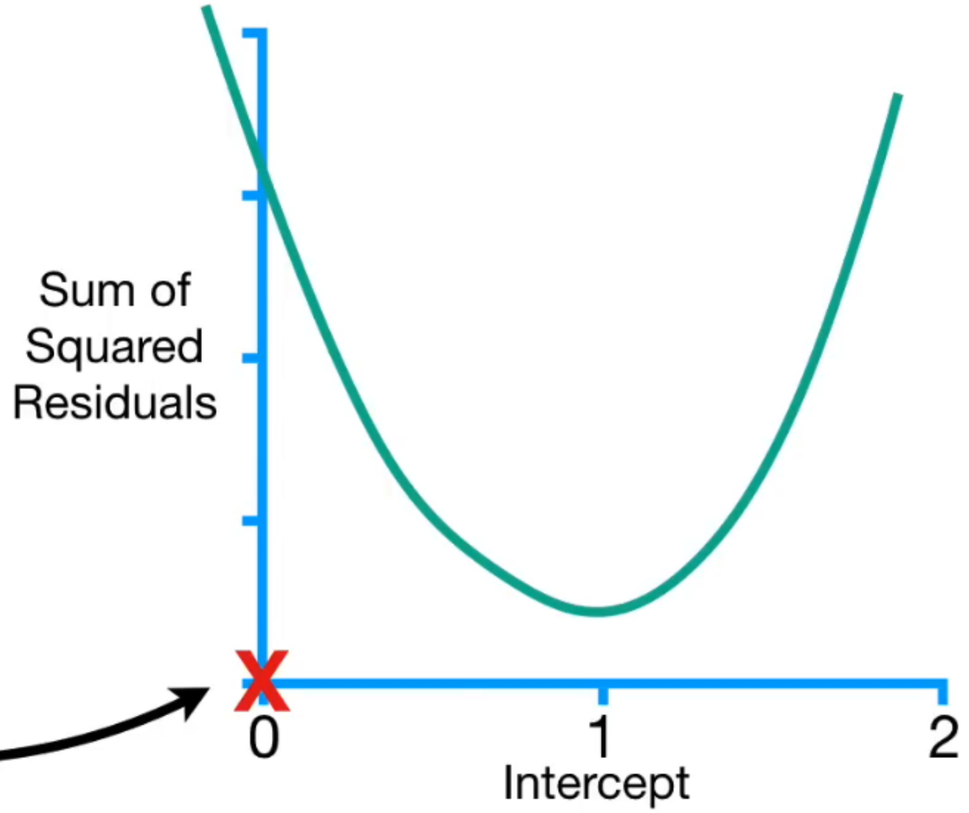
$$-2(\mathbf{1.4} - (\text{intercept} + 0.64 \times \mathbf{0.5})$$

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

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

So we plug 0 into the derivative

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



we will get -5.7

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

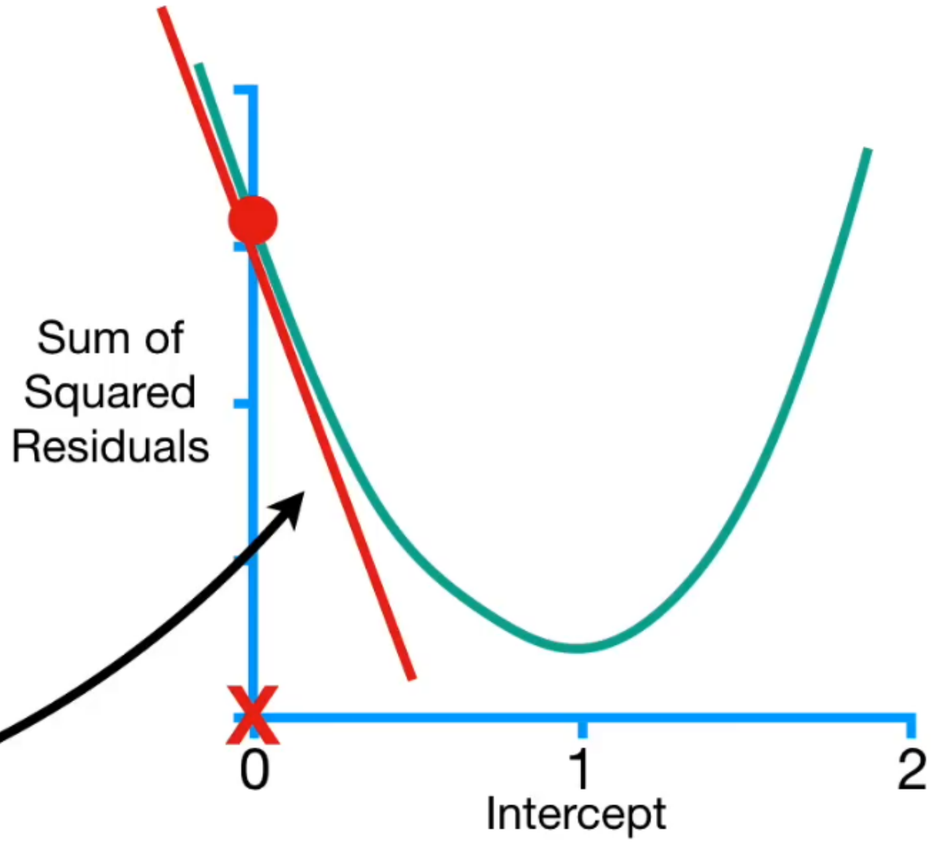
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

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

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

$$= -5.7$$

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



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

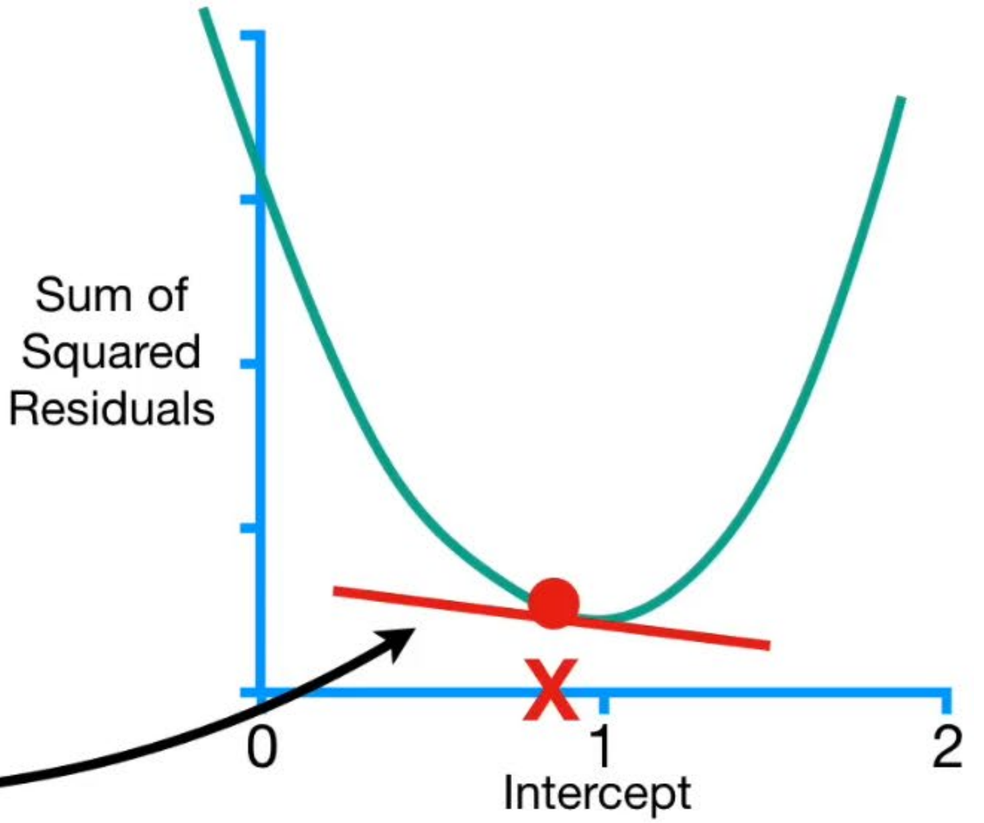
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

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

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

$$= -5.7$$

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



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

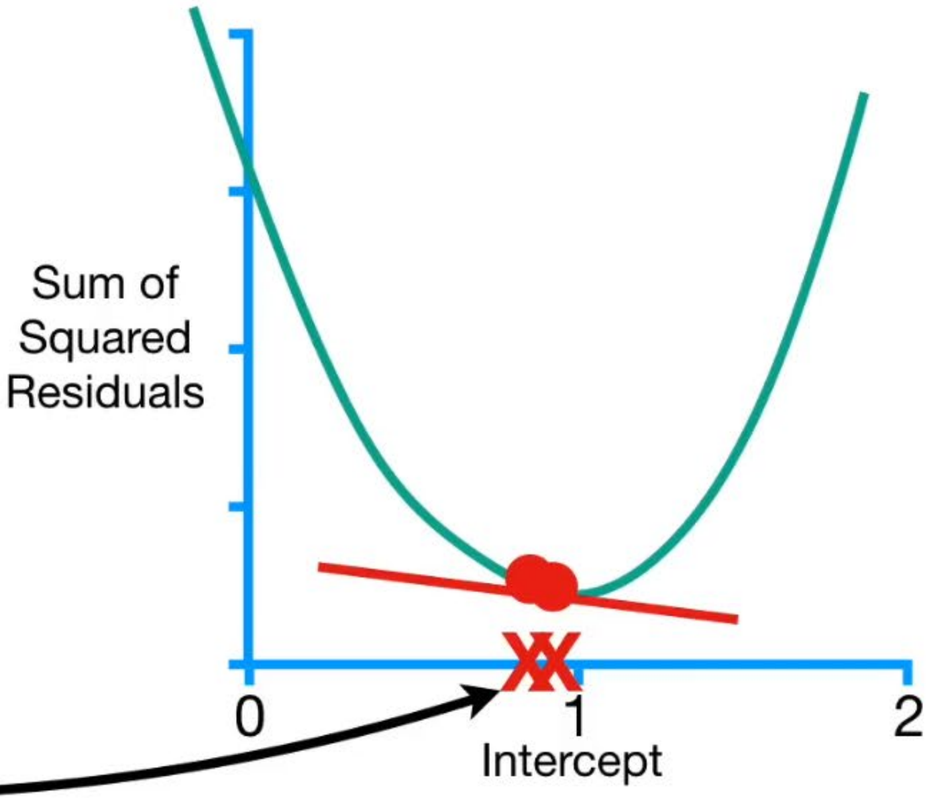
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

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

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

$$= -5.7$$

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



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

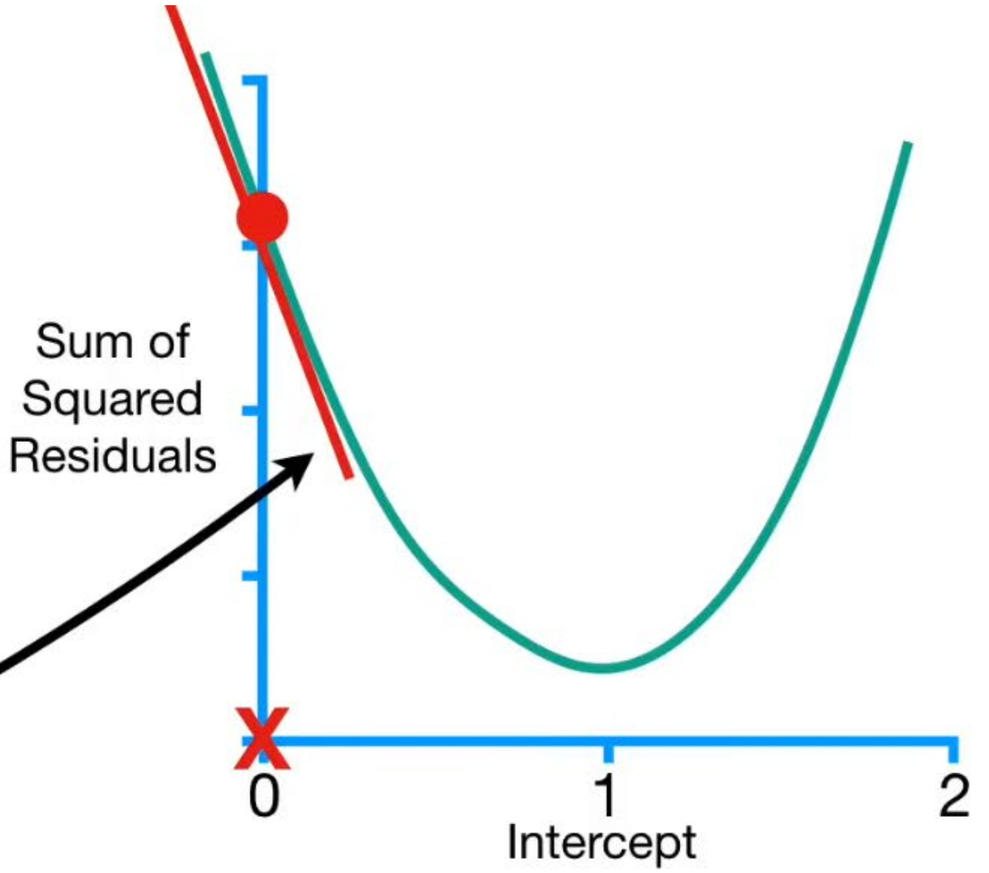
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

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

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

$$= -5.7$$

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



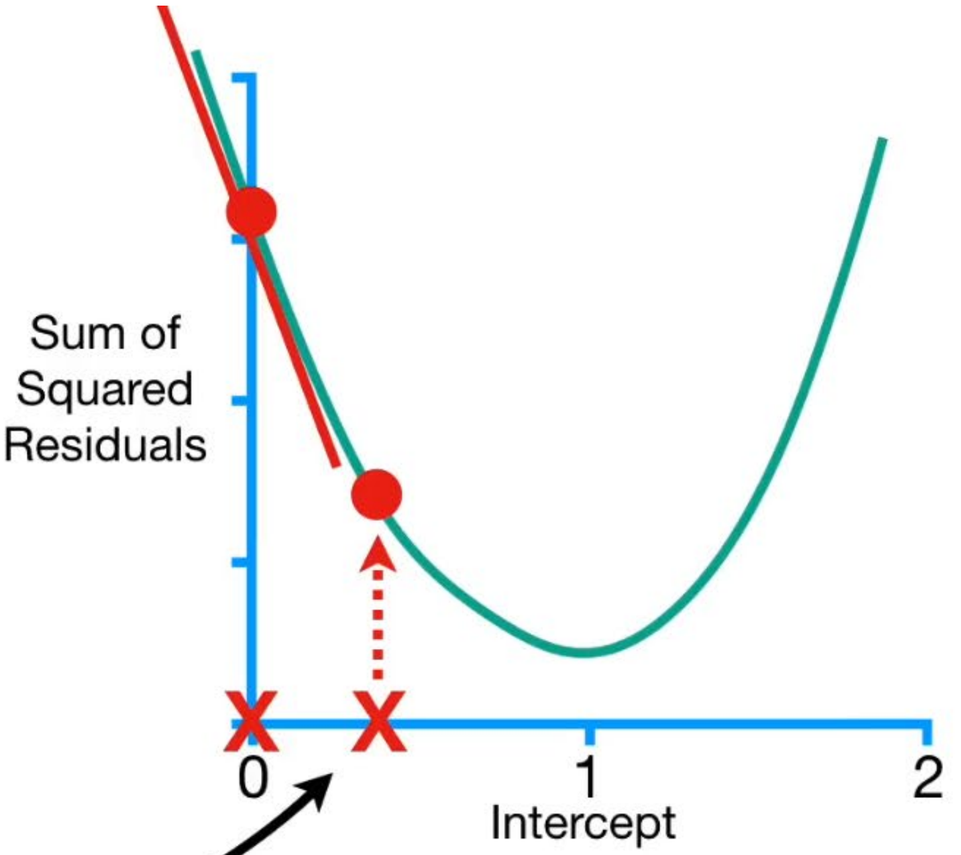
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

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

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

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

$$= -5.7$$



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

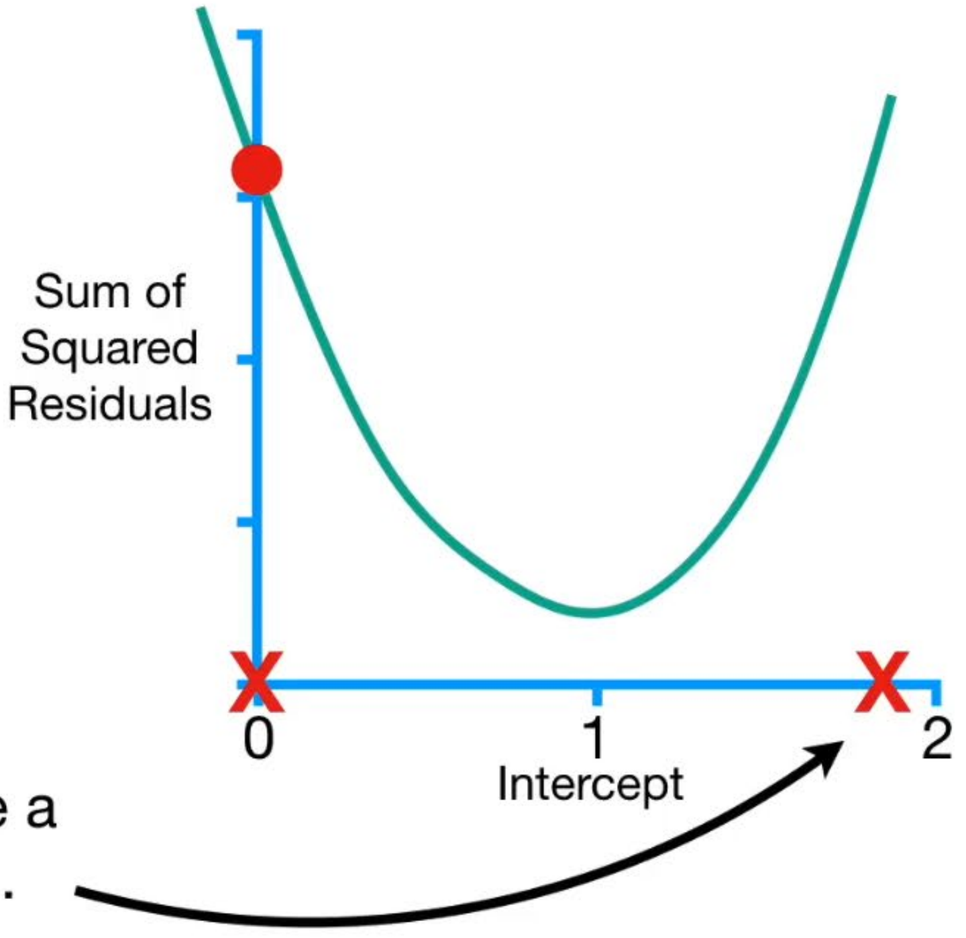
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

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

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

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

$$= -5.7$$



However, if we take a super huge step…

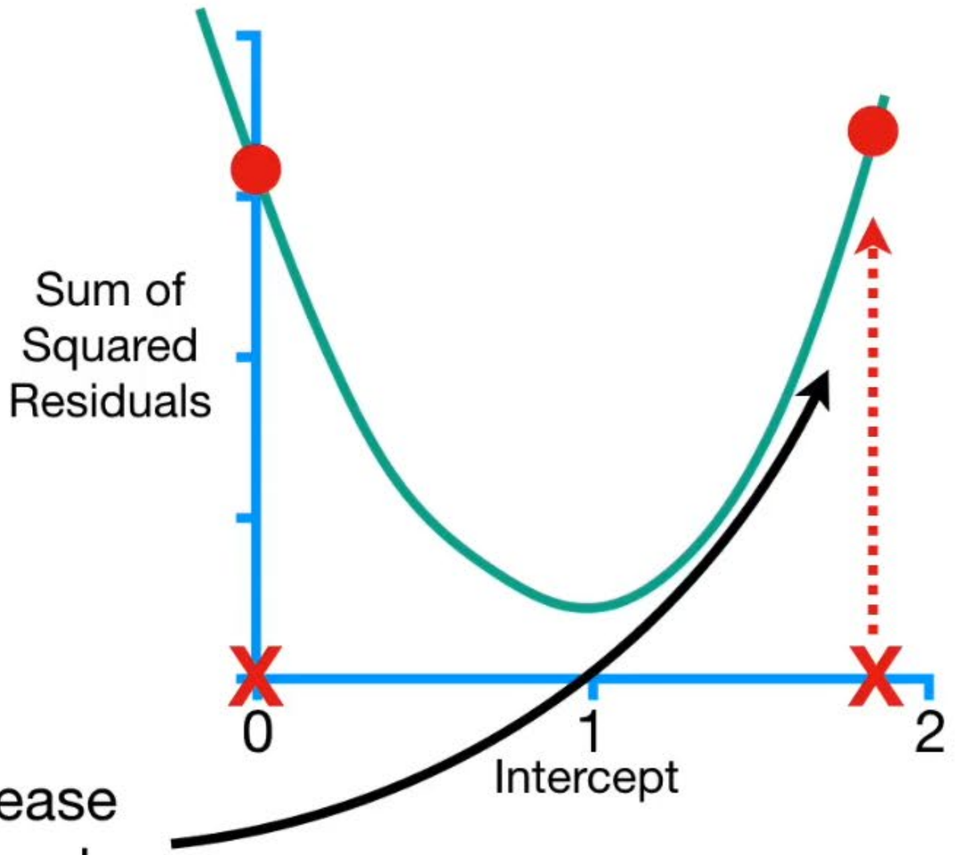
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

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

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

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

$$= -5.7$$



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

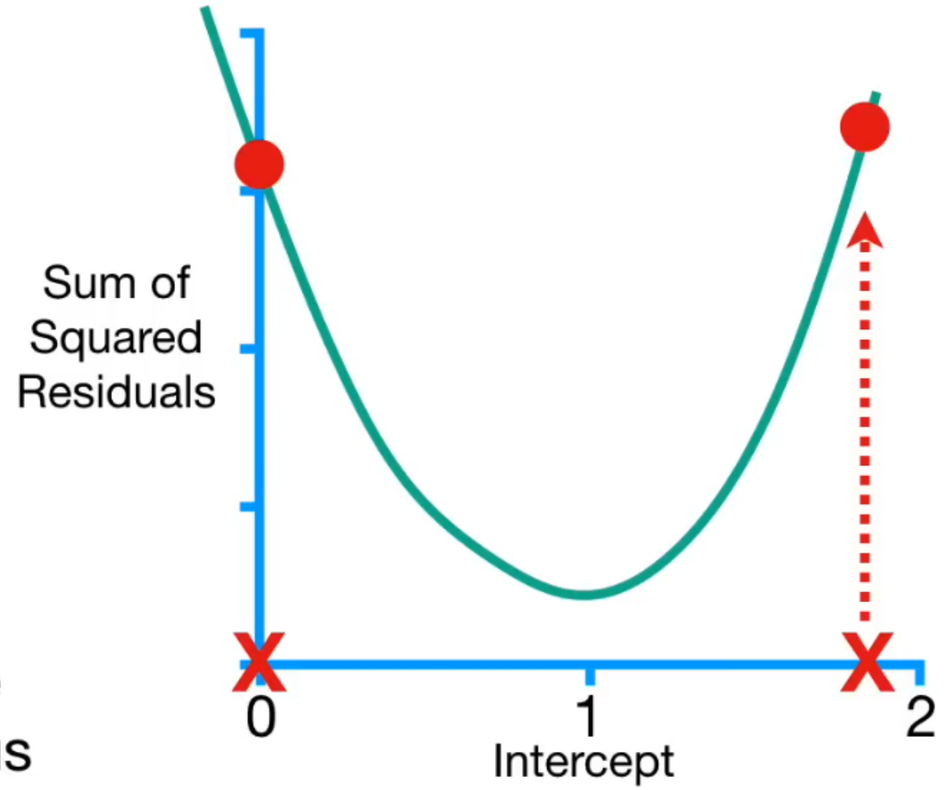
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

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

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

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

$$= -5.7$$



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.

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

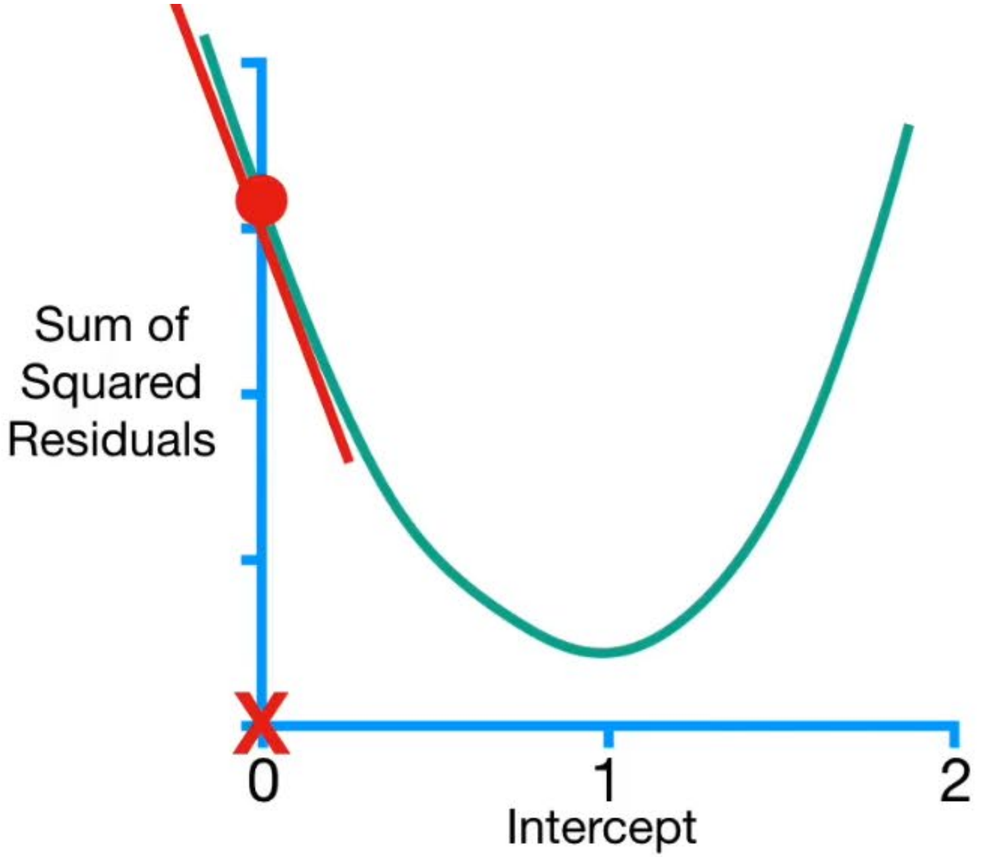
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

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

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

$$= -5.7$$

Step Size = -5.7 x 0.1



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

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

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

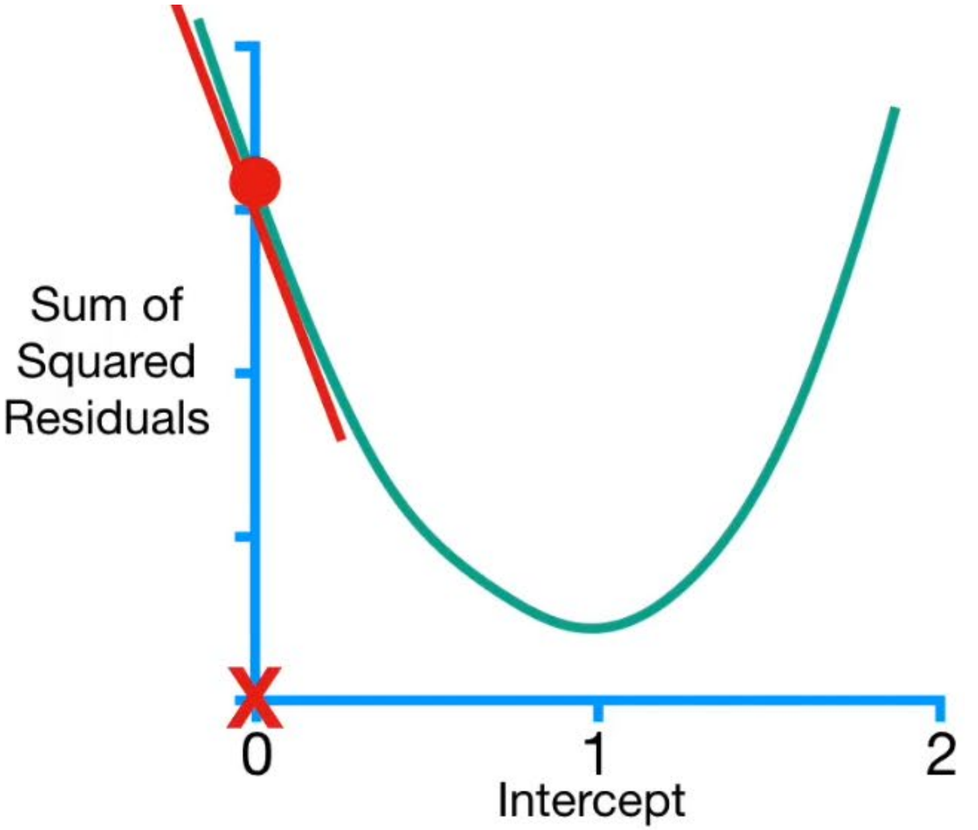
$$+ -2(\mathbf{1.9} - (0 + 0.64 \times \mathbf{2.3}))$$

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

$$= -5.7$$

Step Size = -5.7 x 0.1 = -0.57

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



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

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

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

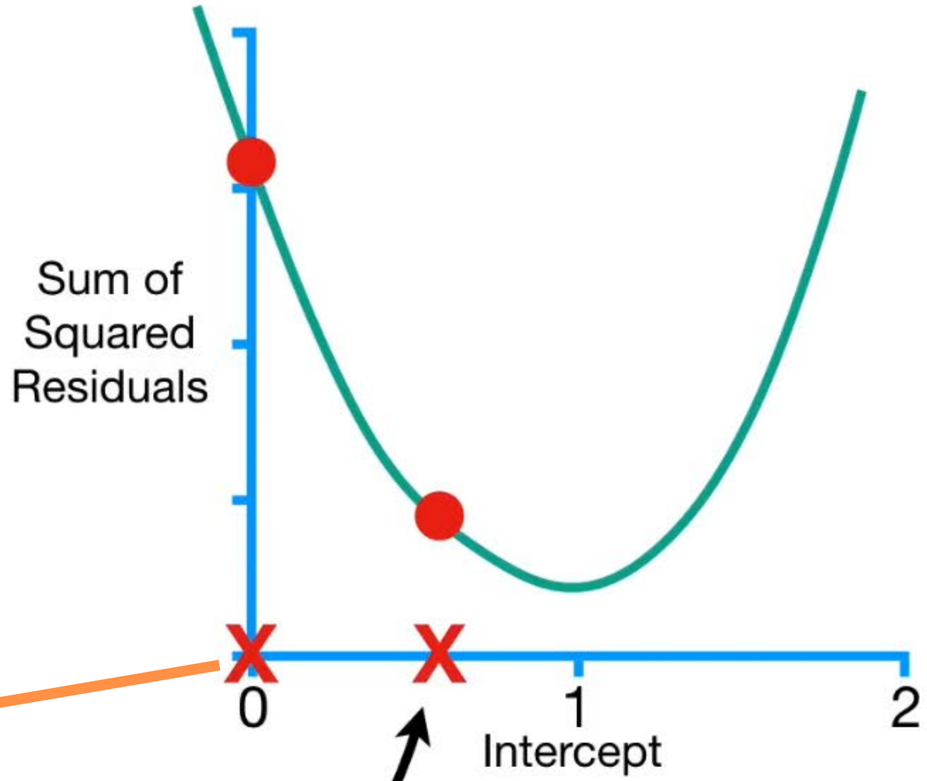
$$-2(\mathbf{1.4} - (0 + 0.64 \times \mathbf{0.5})$$

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

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

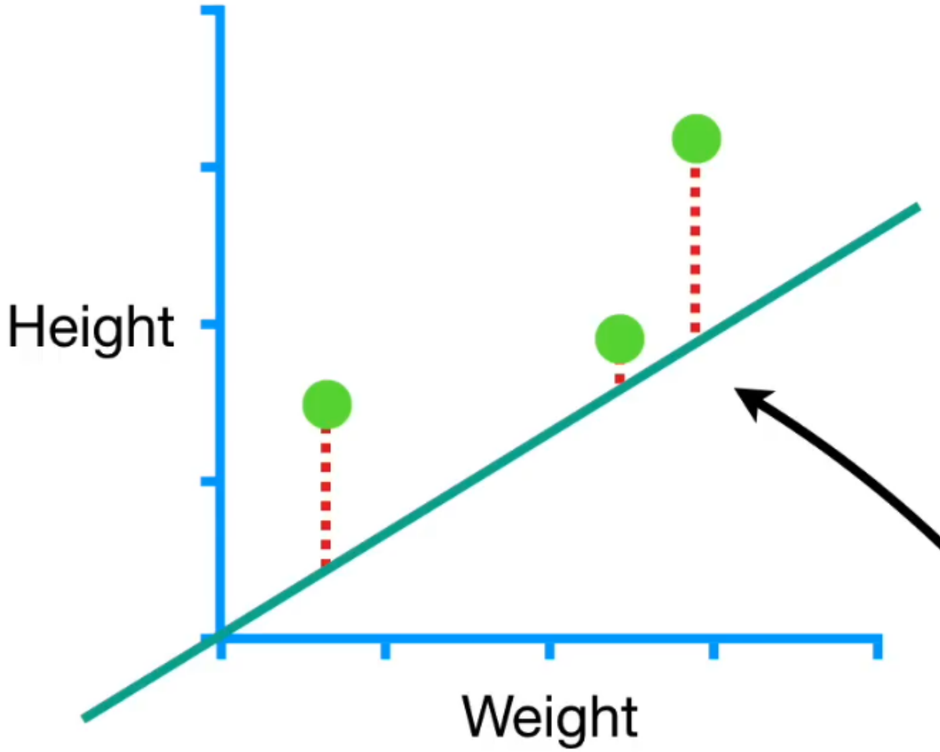
$$= -5.7$$

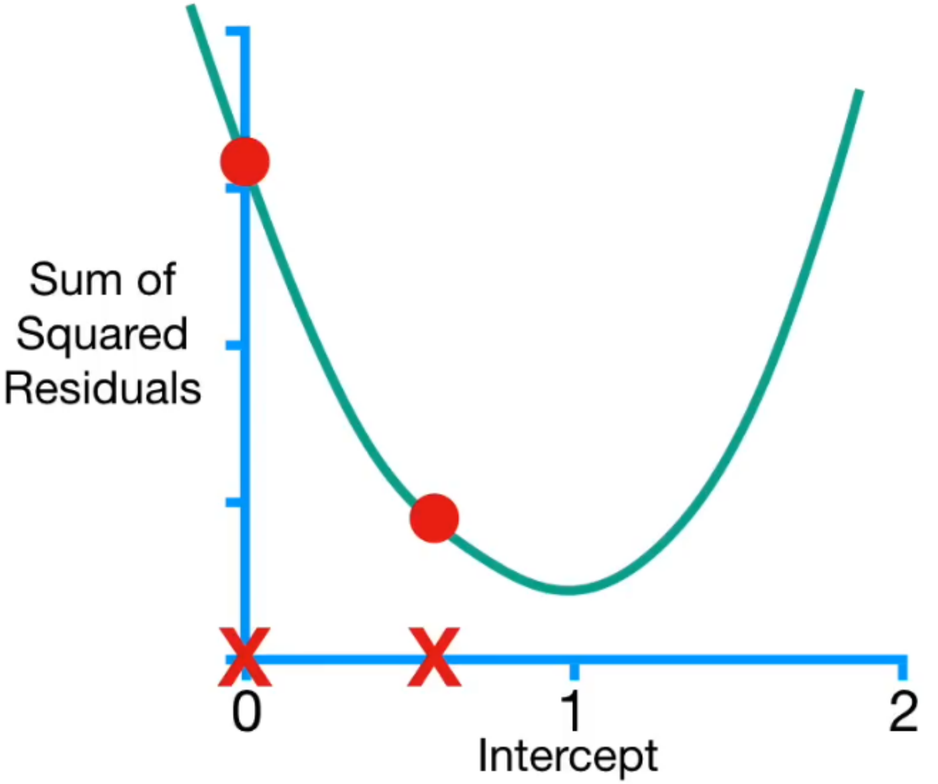
Step Size = -5.7 × 0.1 = -0.57



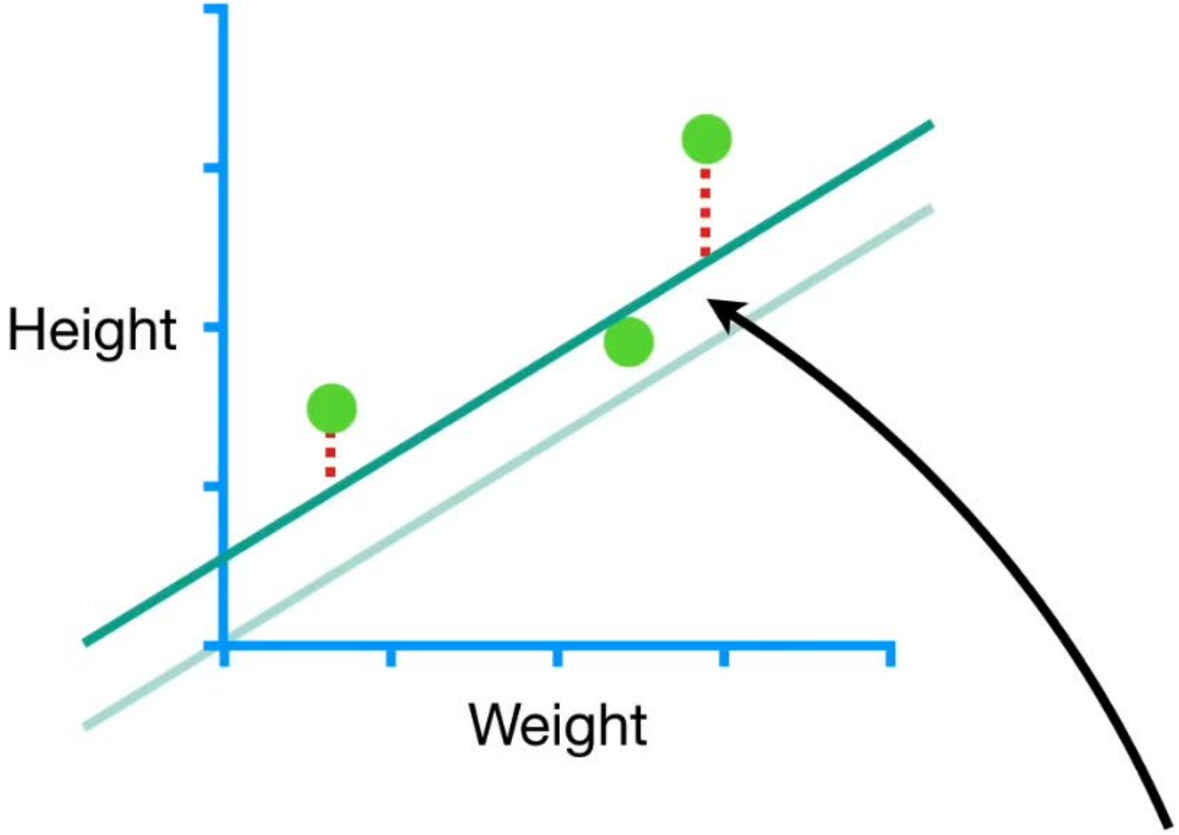
New Intercept = 0 - (-0.57) = **0.57**

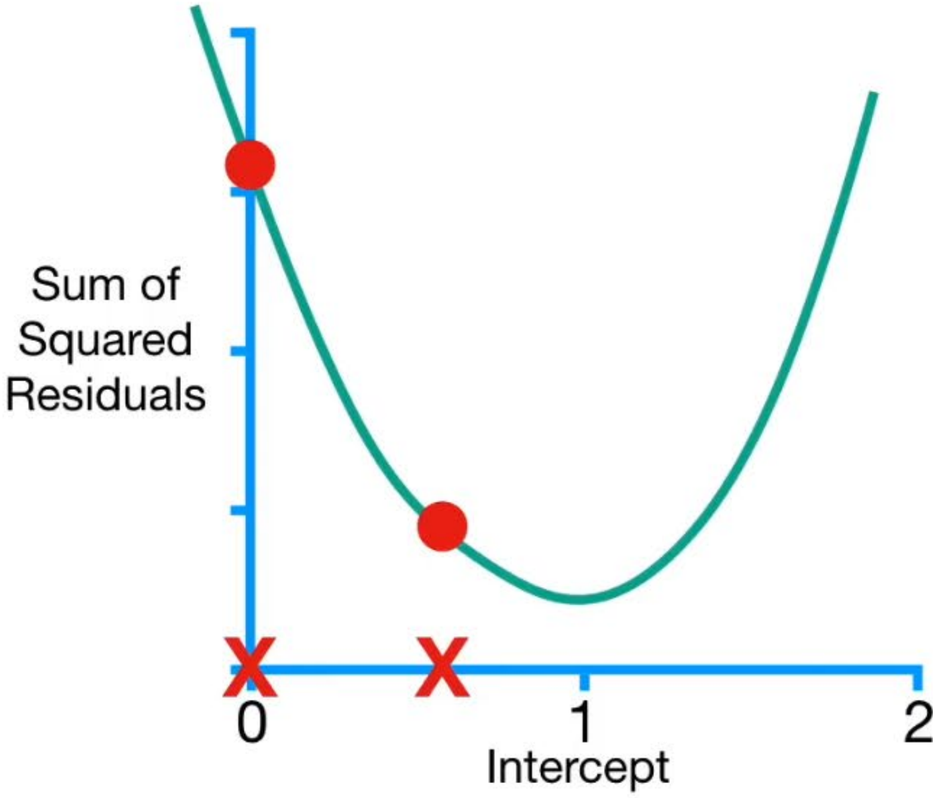
…and the the **New Intercept** = **0.57**.





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**.

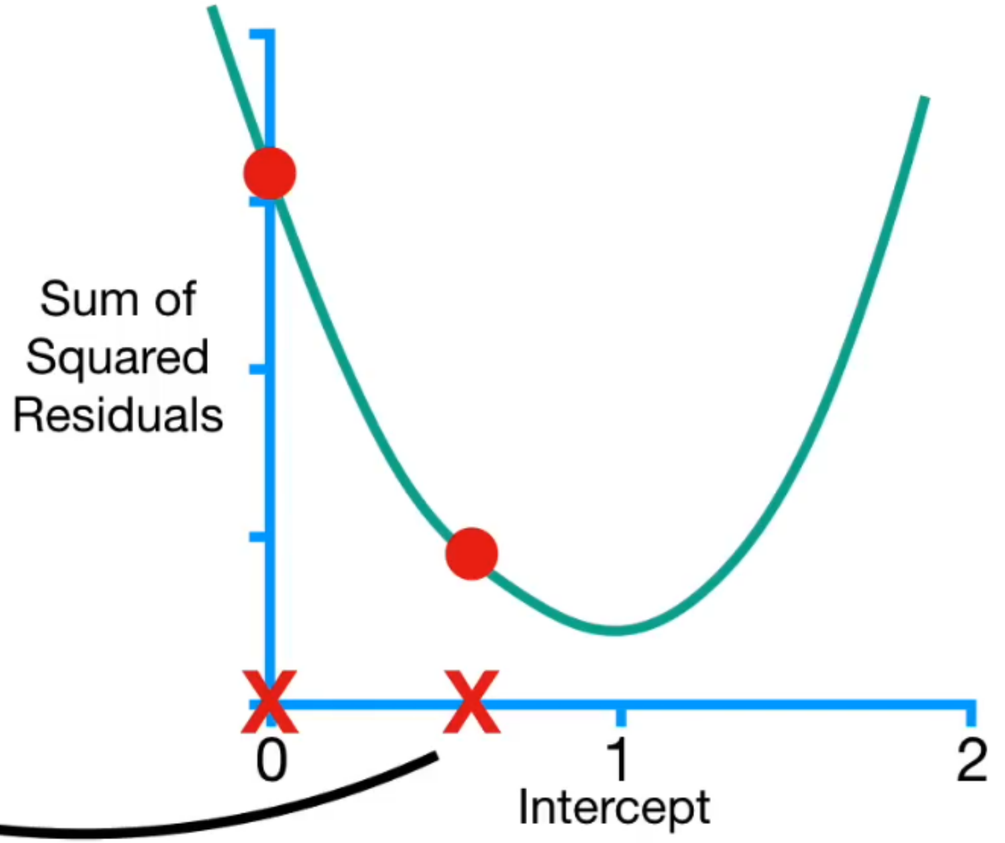
$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

-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 =

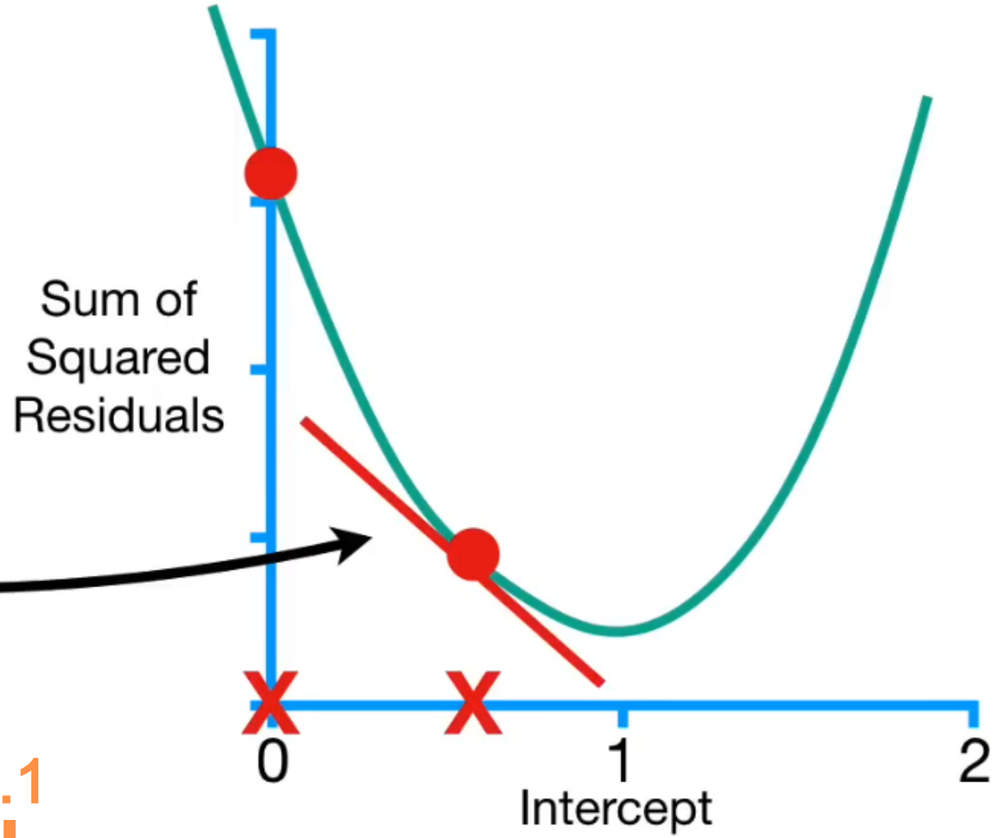
$$-2(\mathbf{1.4} - (0.57 + 0.64 \times \mathbf{0.5})$$

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

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

$$= \boxed{-2.3}$$

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



0.1

Step Size = Slope × Learning Rate

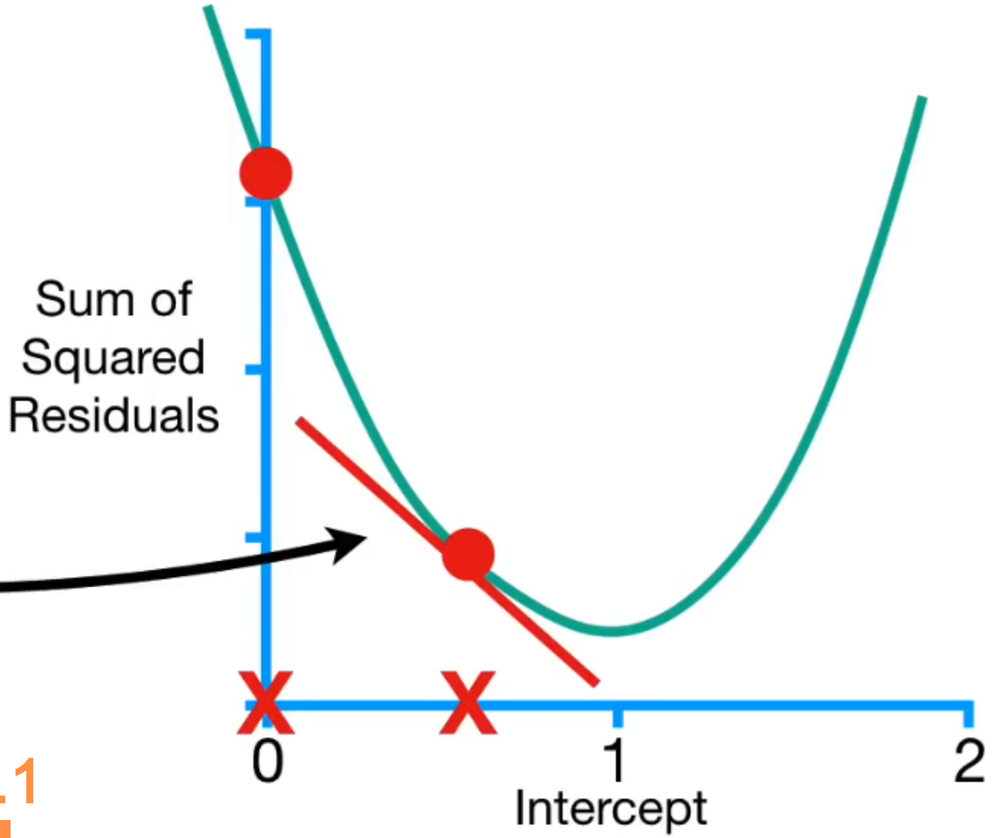
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

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

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

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

$$= -2.3$$



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

0.1

Step Size = Slope × Learning Rate

Step Size = -2.3 × 0.1 = -0.23

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

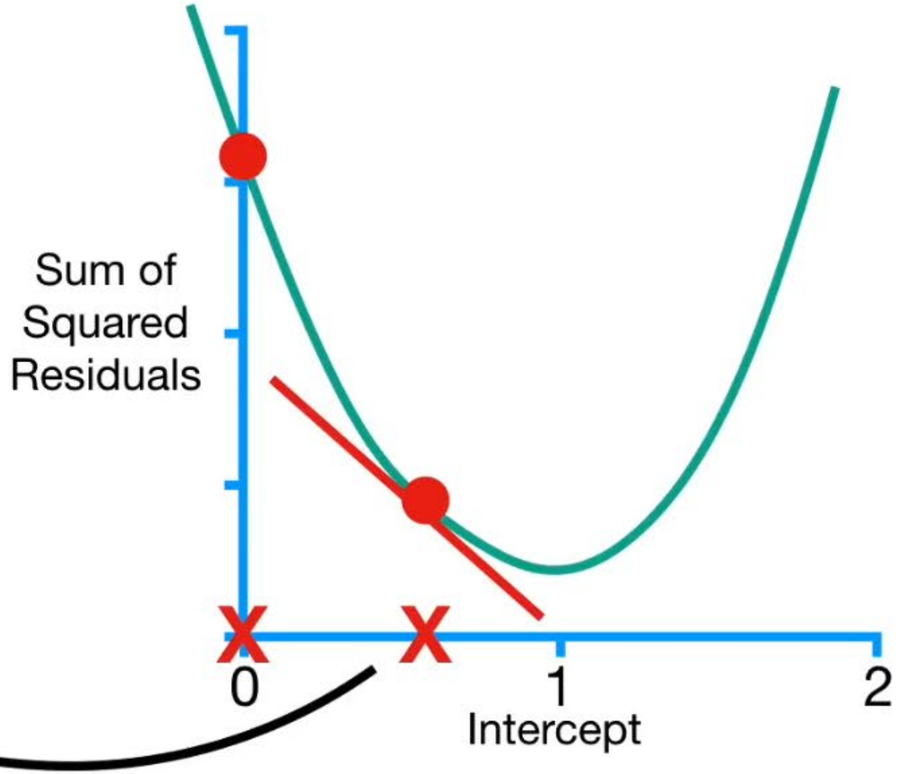
$$-2(\mathbf{1.4} - (0.57 + 0.64 \times \mathbf{0.5})$$

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

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

$$= -2.3$$

Step Size = -2.3 × 0.1 = **-0.23**



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

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

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

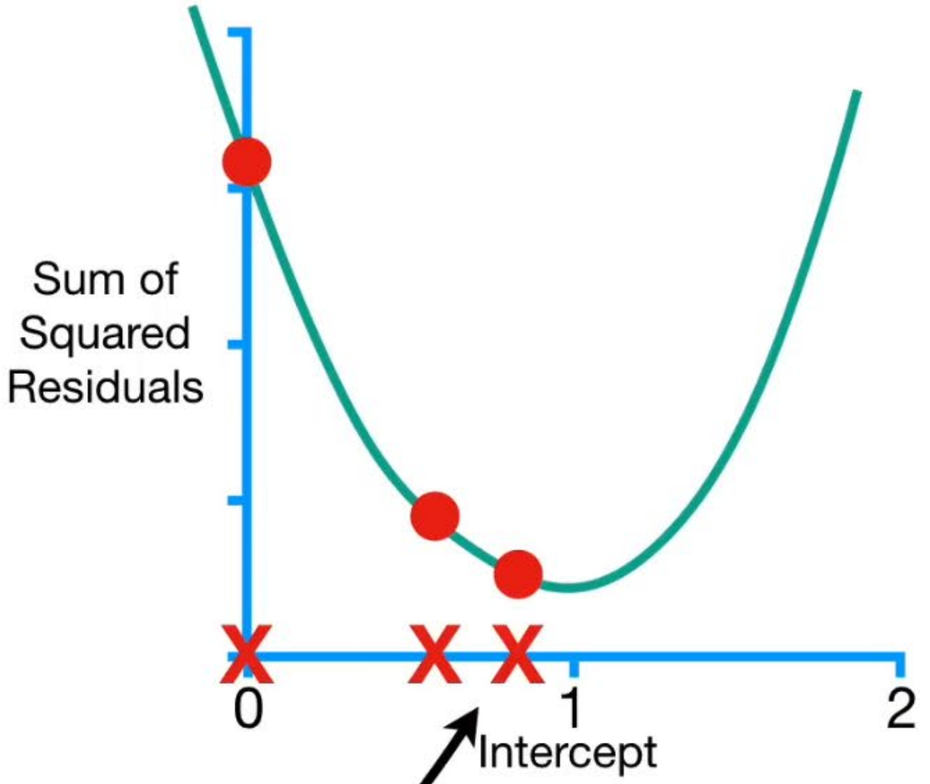
$$-2(\mathbf{1.4} - (0.57 + 0.64 \times \mathbf{0.5}))$$

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

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

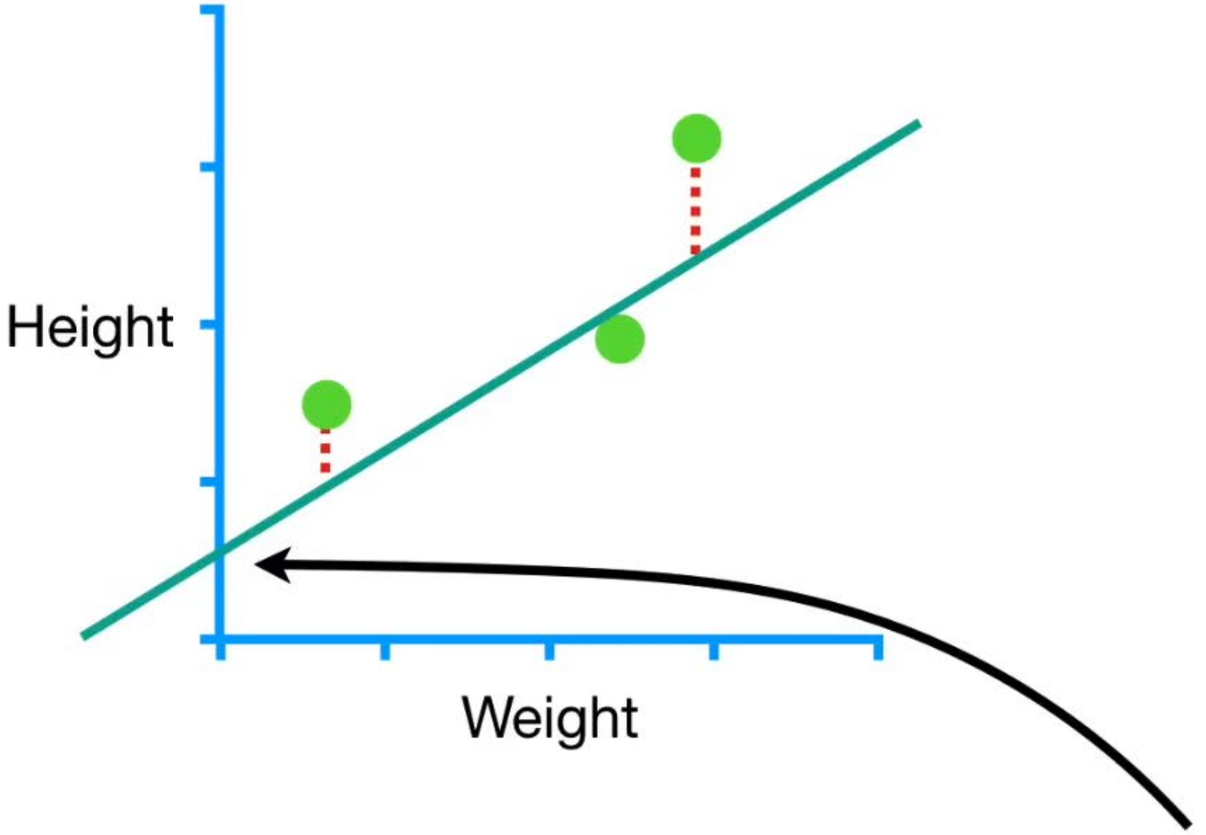
$$= -2.3$$

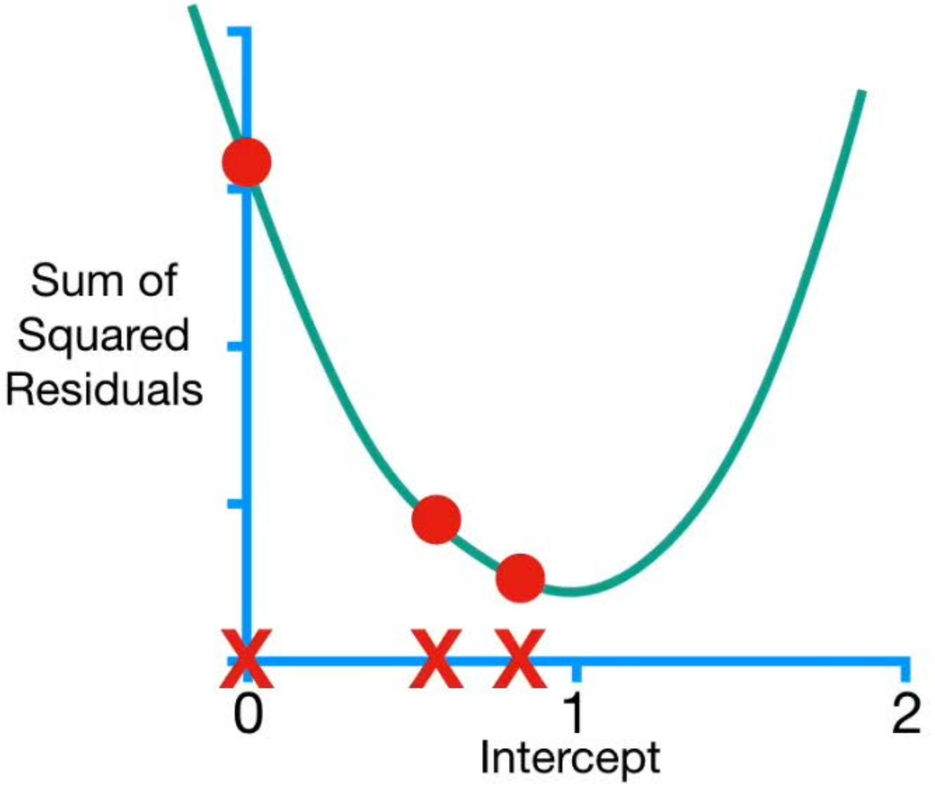
Step Size = -2.3 × 0.1 = **-0.23**



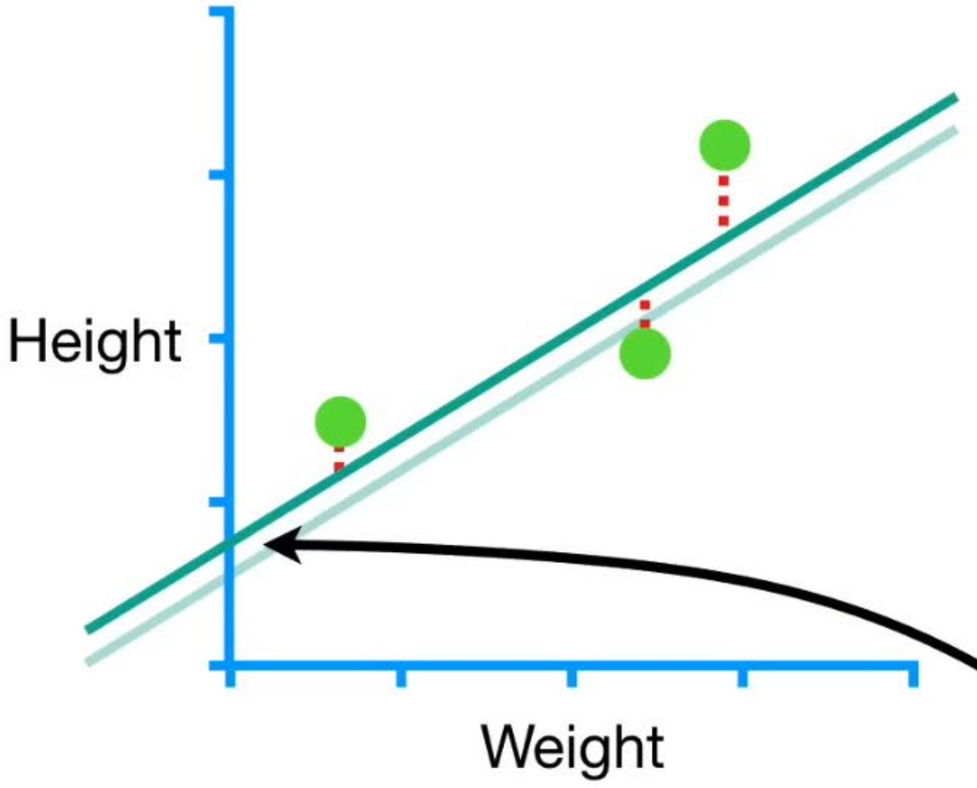
New Intercept = 0.57 - (-0.23) = **0.8**

…and the **New Intercept** = **0.8**

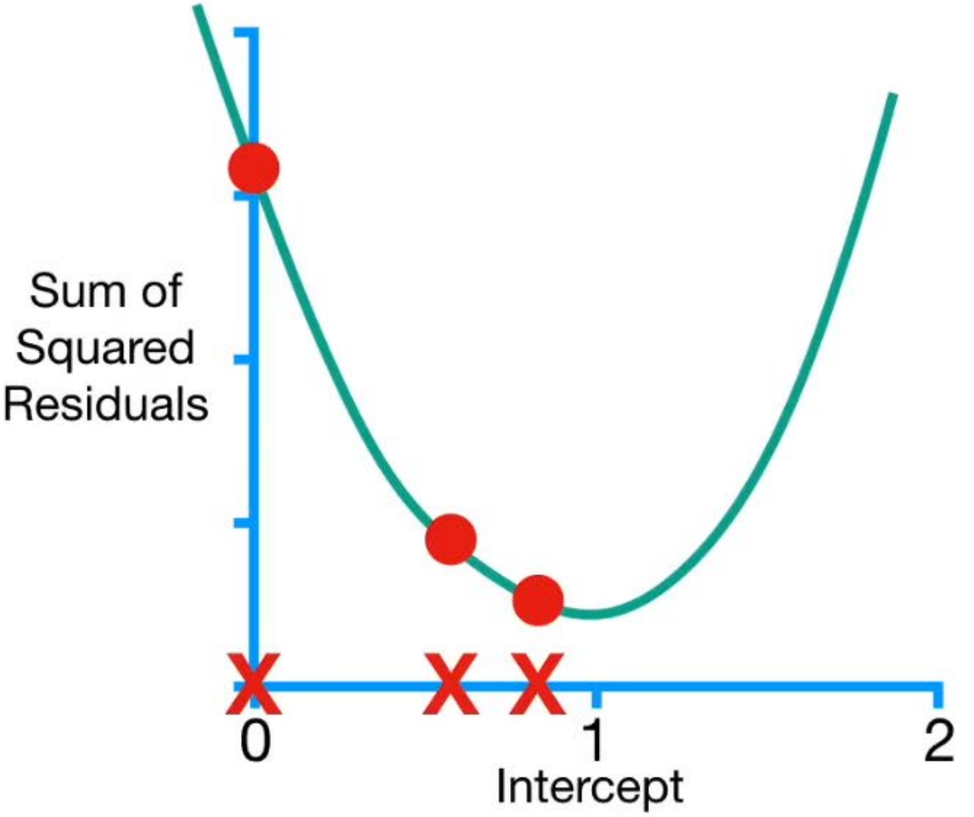




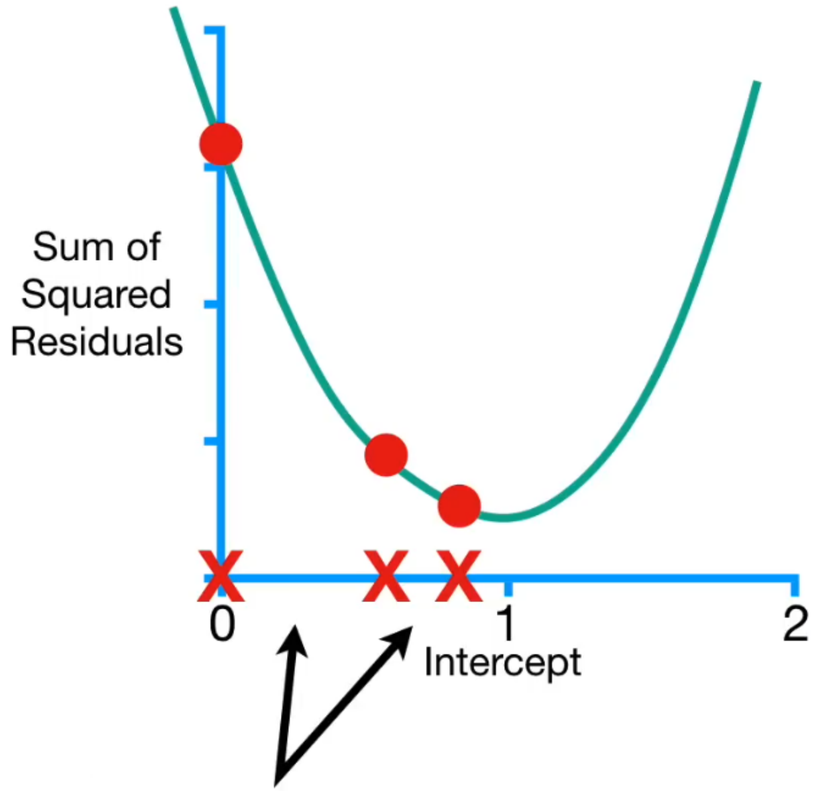
Now we can compare the residuals when the **Intercept** = **0.57**…



…to when the
Intercept = 0.8



Overall, the Sum of the Squared Residuals is getting smaller.



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

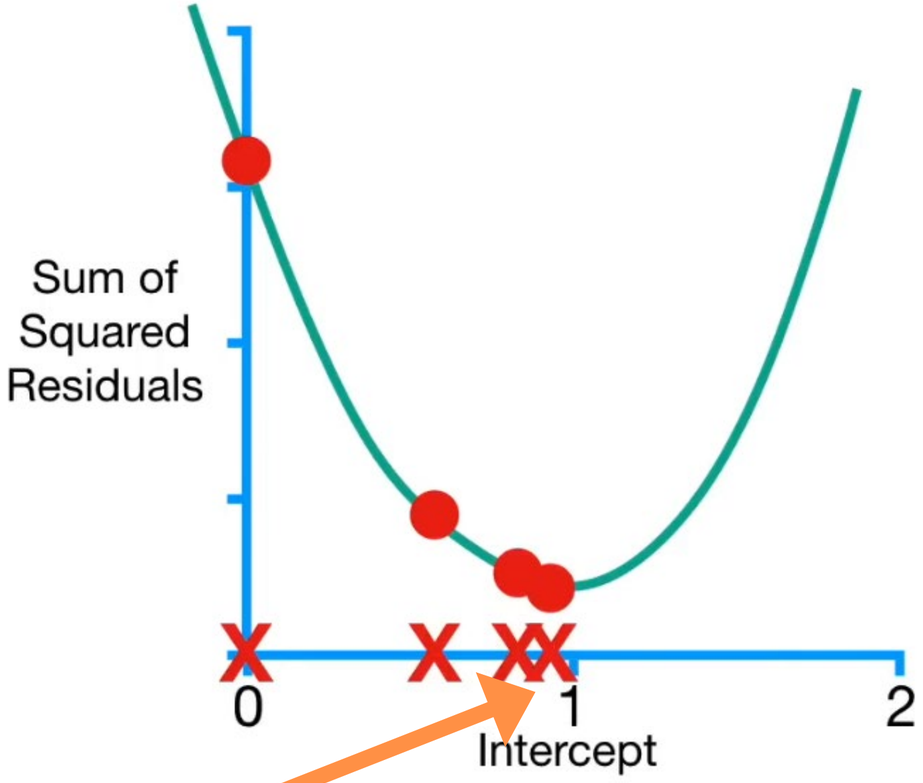
$$= \boxed{\mathbf{-0.9}}$$

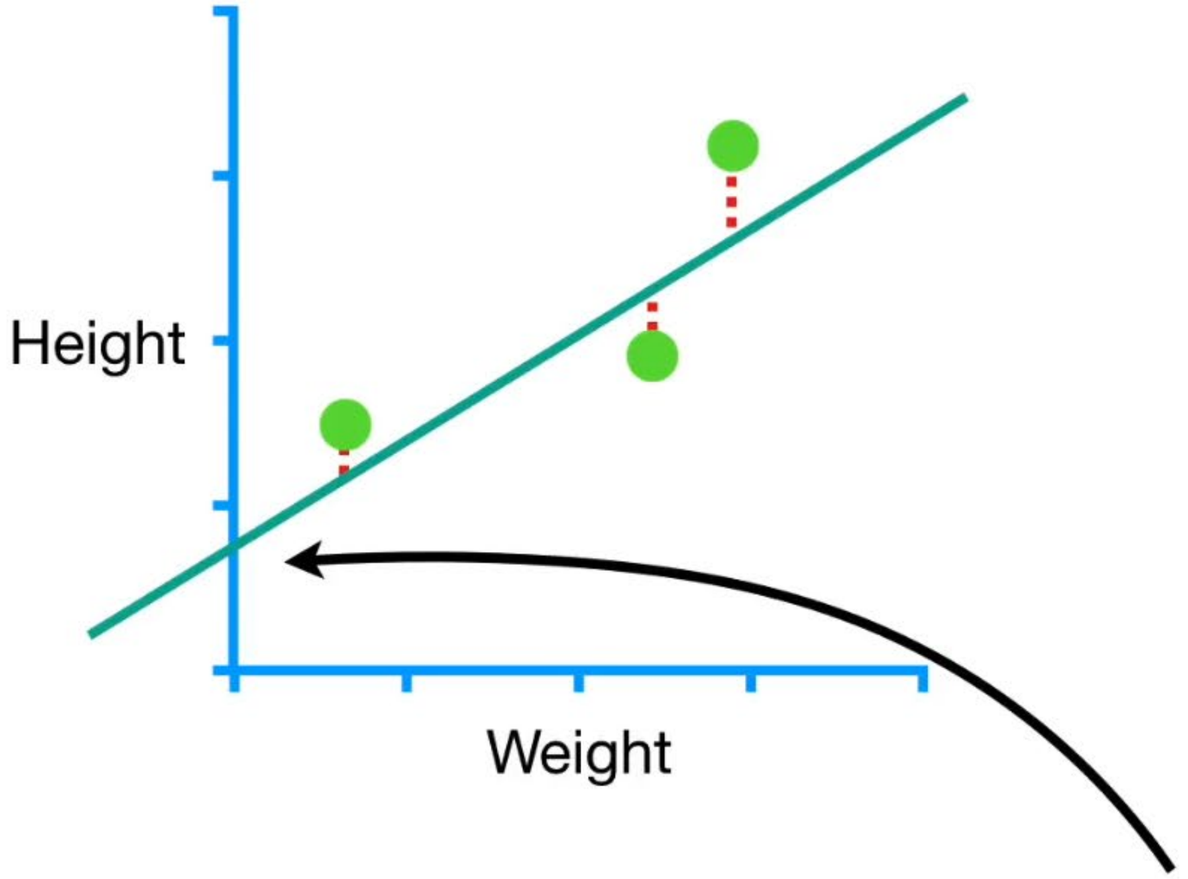
Step Size = **Slope** × **Learning Rate**

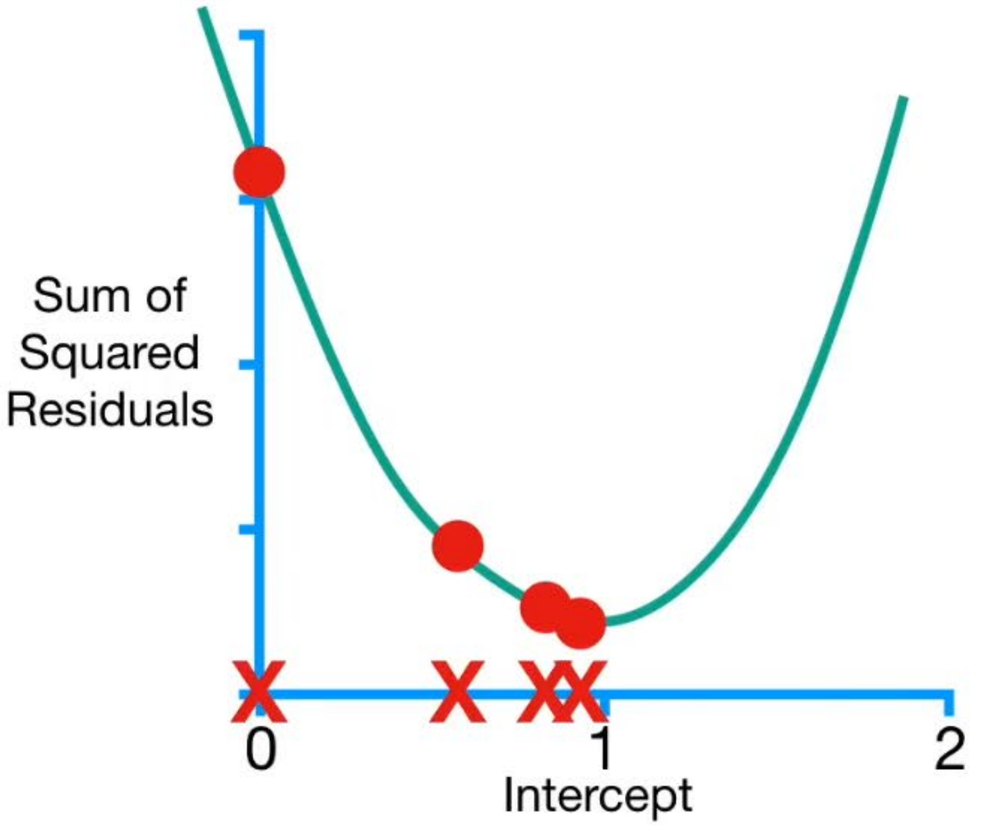
Step Size = -0.9 × 0.1 = **-0.09**

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

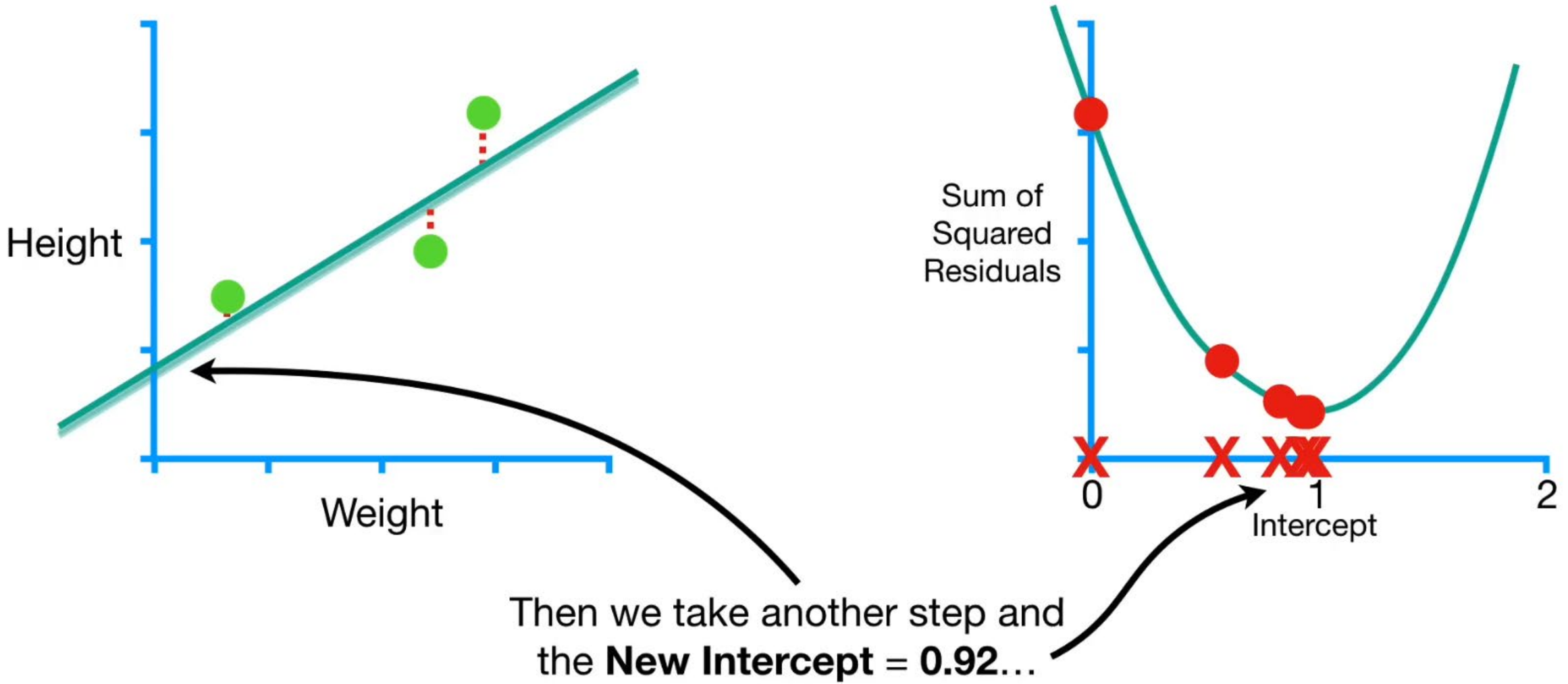
New Intercept = 0.8 - (-0.09) = **0.89**

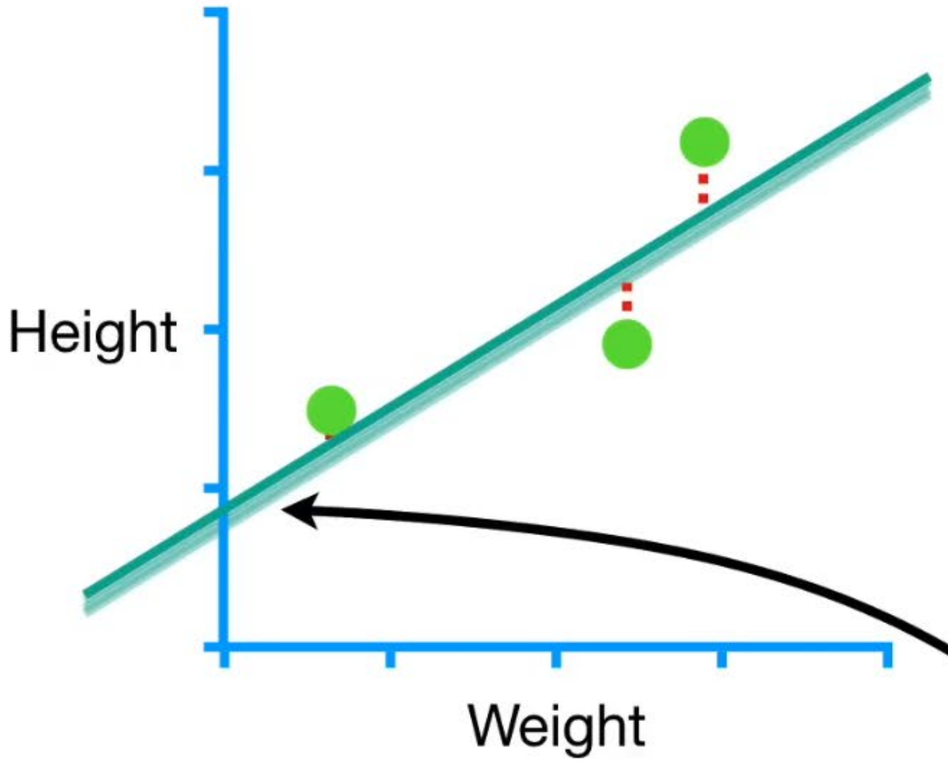


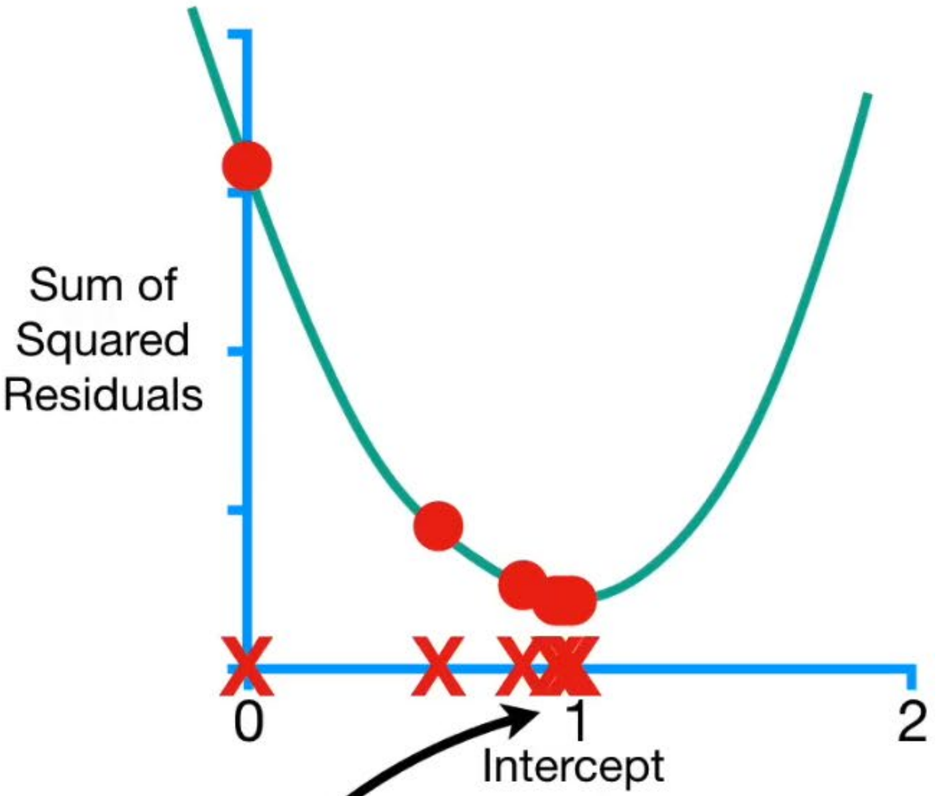




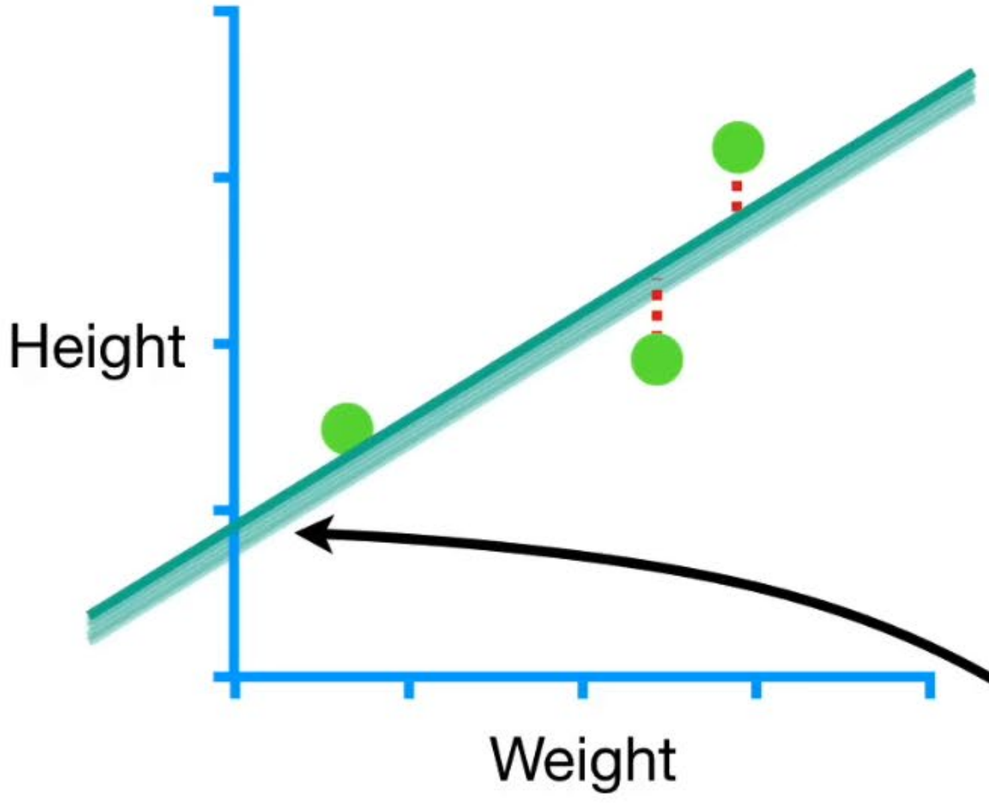
Now we increase the **Intercept** from **0.8**…

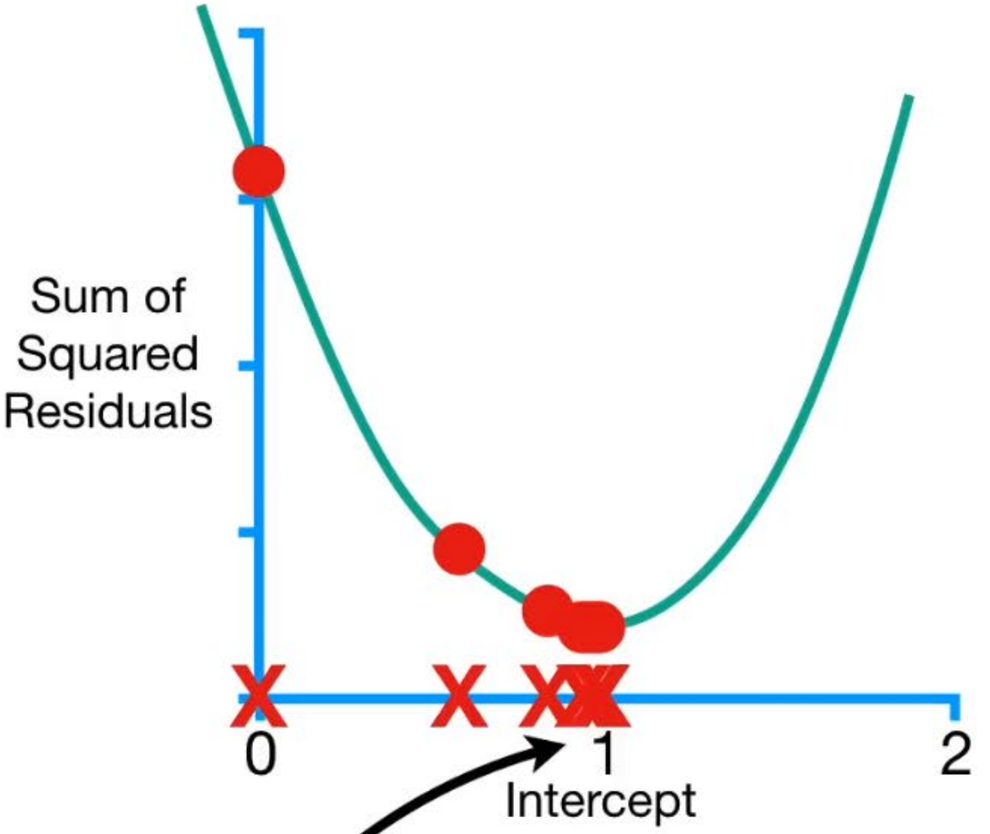

Height
Weight
Sum of
Squared
Residuals
0
1
2
Intercept
Then we take another step and
the **New Intercept** = **0.92**…



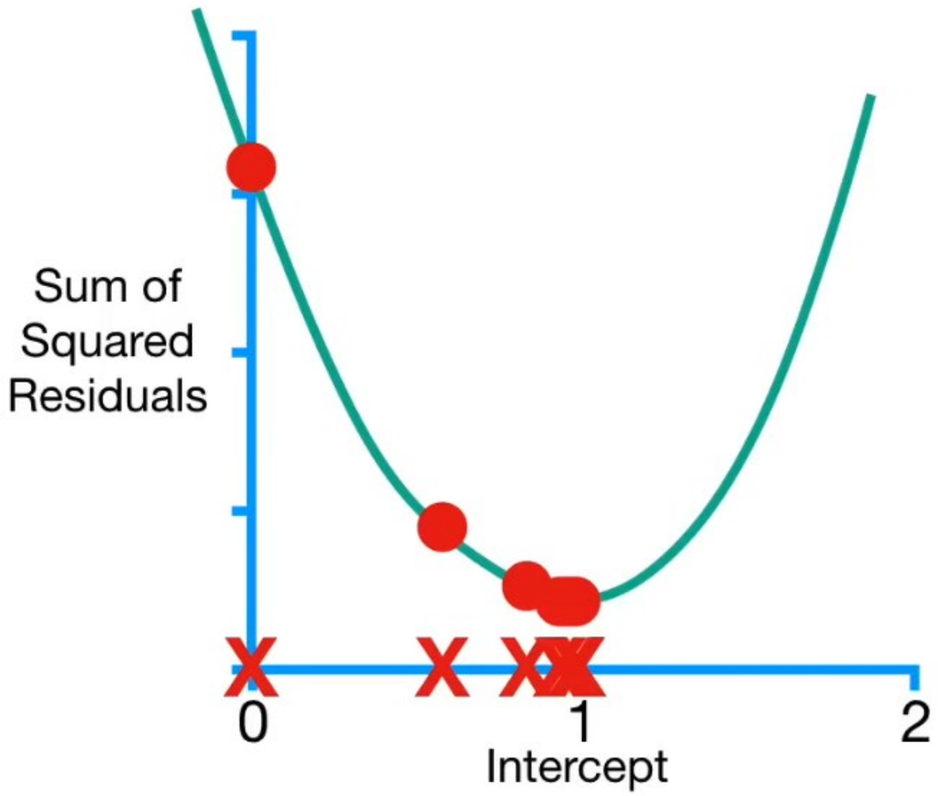


…and then we take another step and the **New Intercept = 0.94**…





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



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

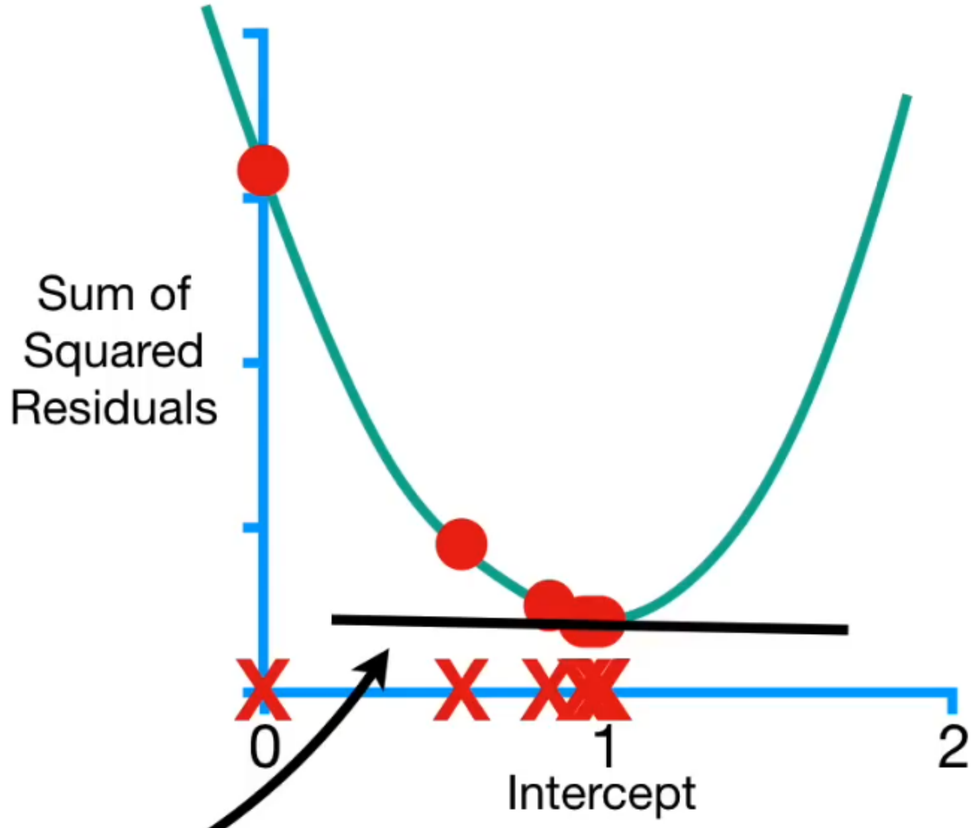
After 6 steps, **the Gradient Descent** estimate for the **Intercept** is **0.95**.

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?

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

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



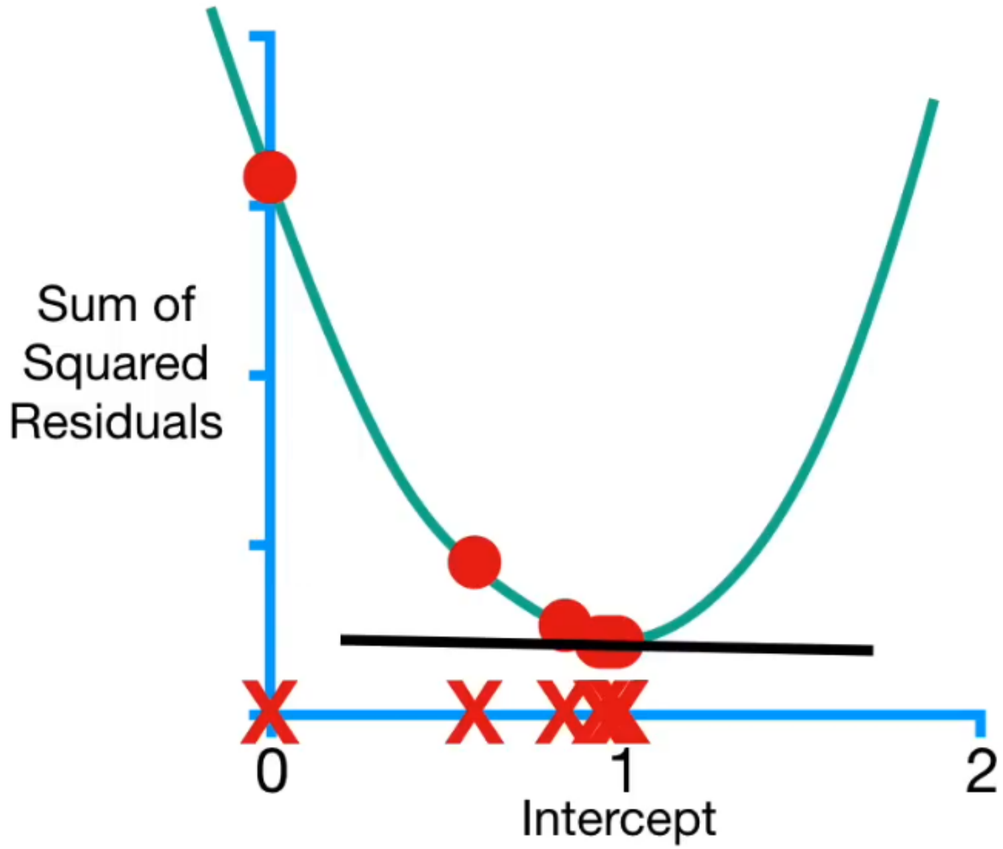
In practice, the Minimum Step Size = 0.001 or smaller.

So if this slope = 0.009

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

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

Step Size = **0.009** x **0.1** = 0.0009



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

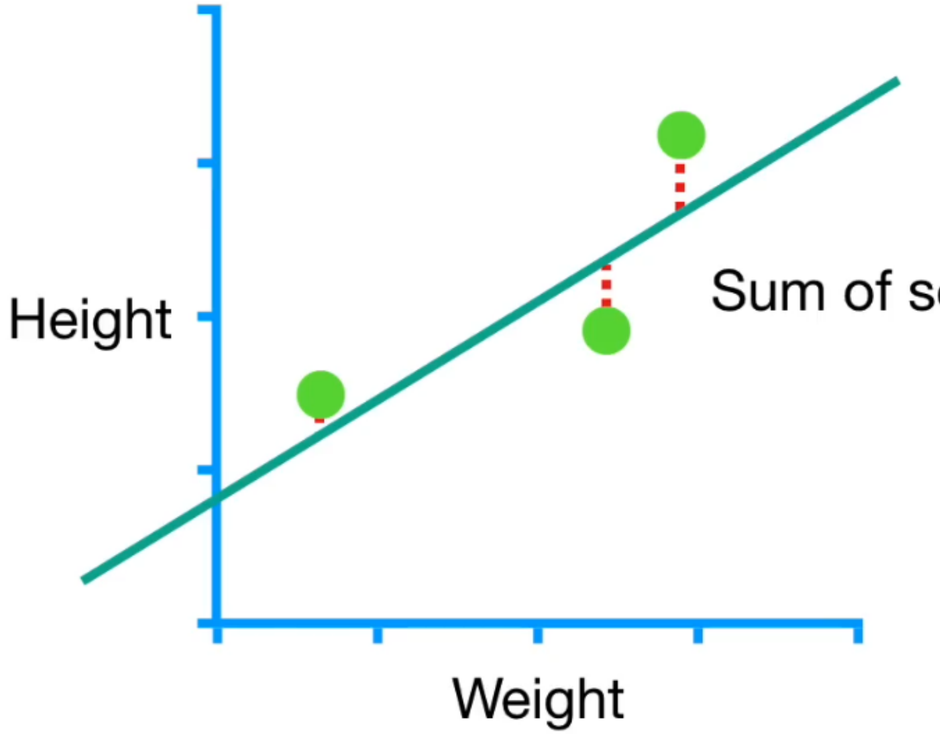
That said, Gradient Descent also includes a limit on the number of steps it will take before giving up.

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

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

OK, let’s review what we’ve learned so far...

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…

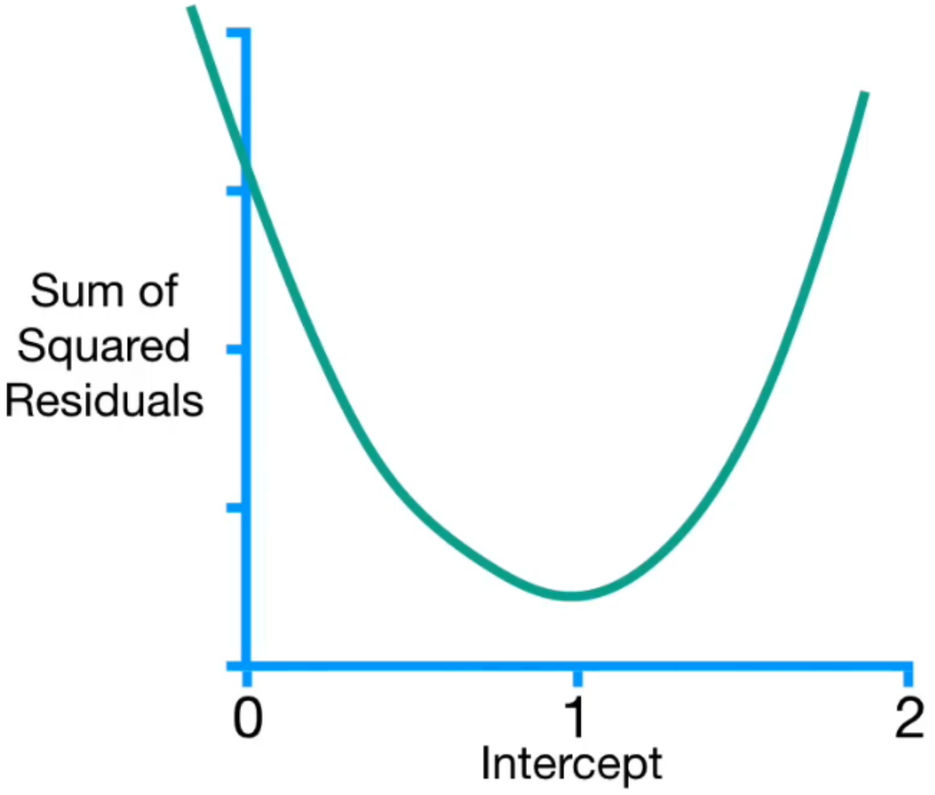


Sum of squared residuals = (**1.4** - (intercept + 0.64 × **0.5**))^2

+ (**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**…



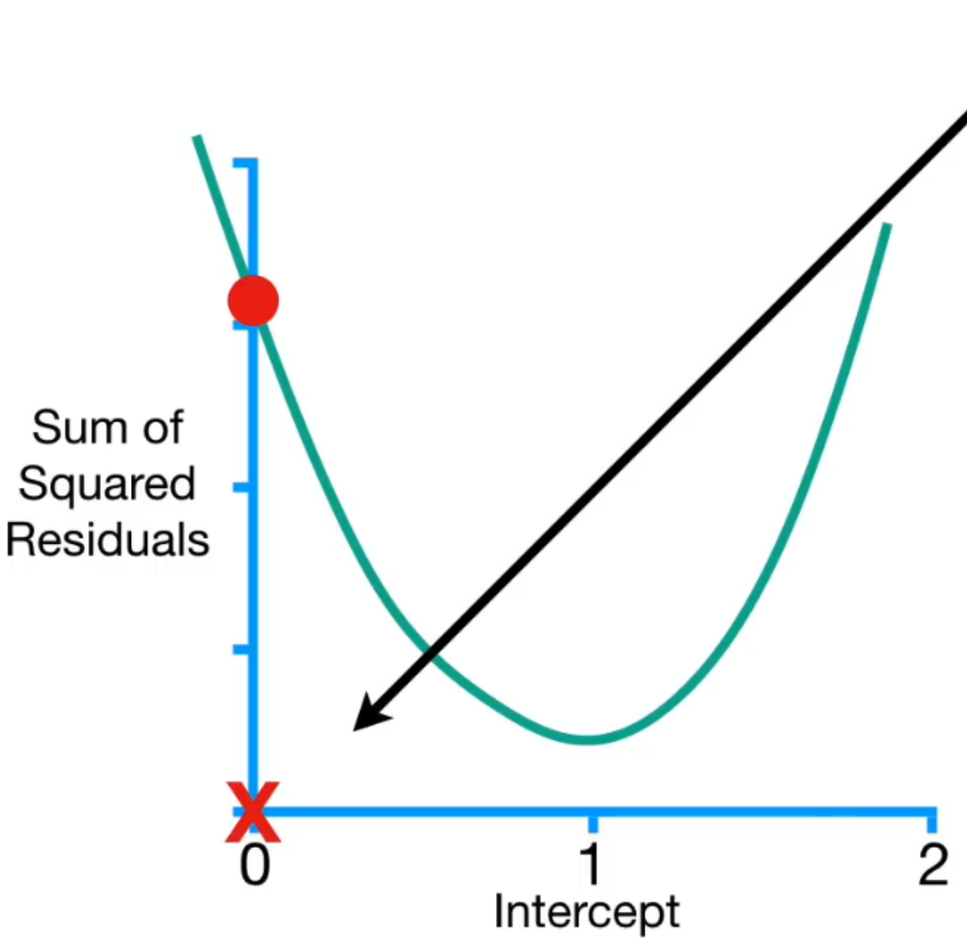
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} =$$

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

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

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

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



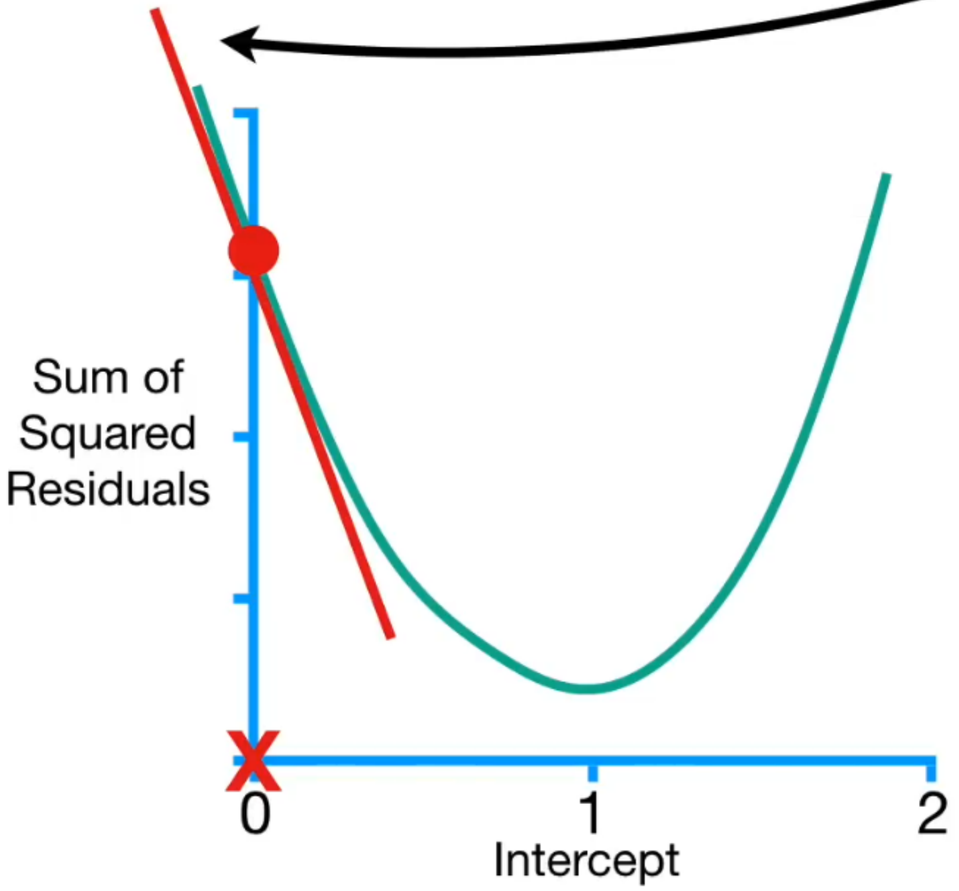
$$\frac{d}{d\ intercept}\ \text{Sum of squared residuals} =$$

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

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

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

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



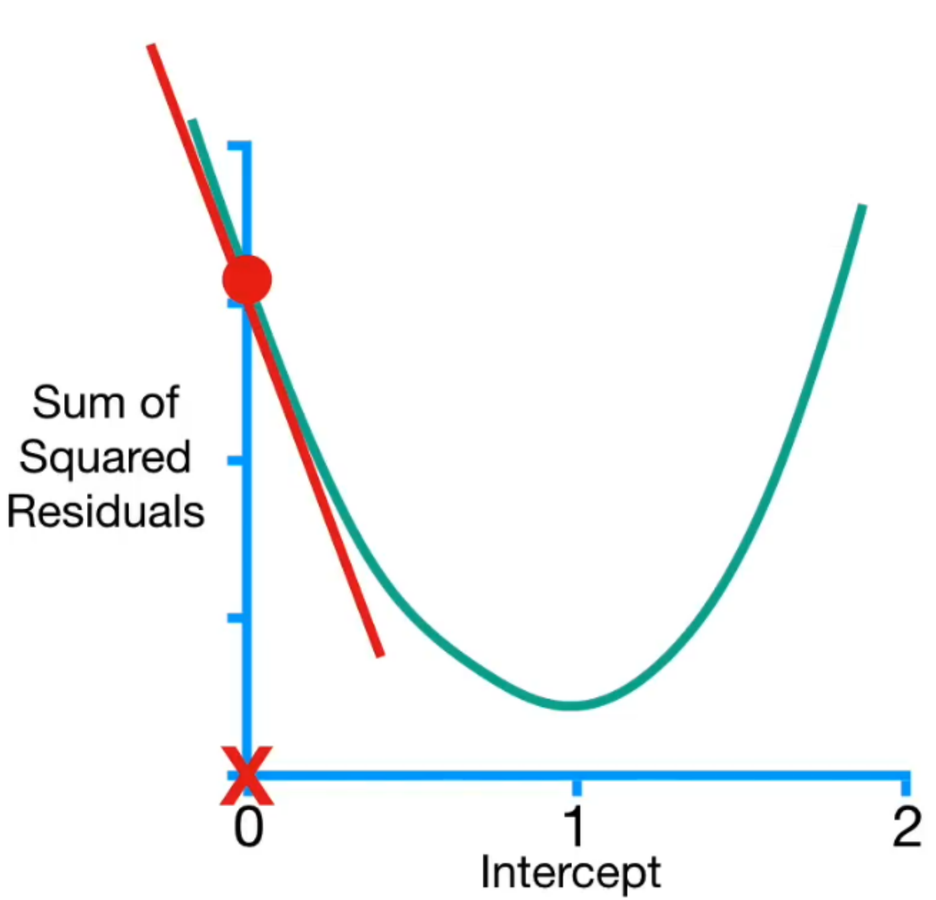
$$\frac{d}{d\ intercept}\ \text{Sum of squared residuals} =$$

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

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

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

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

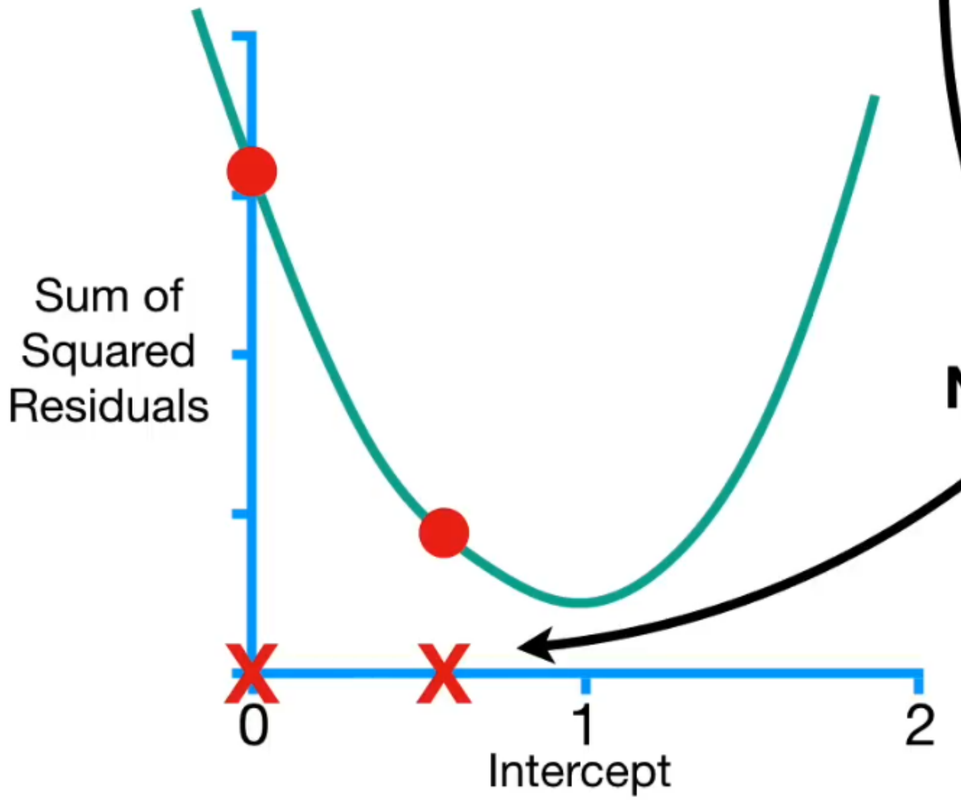


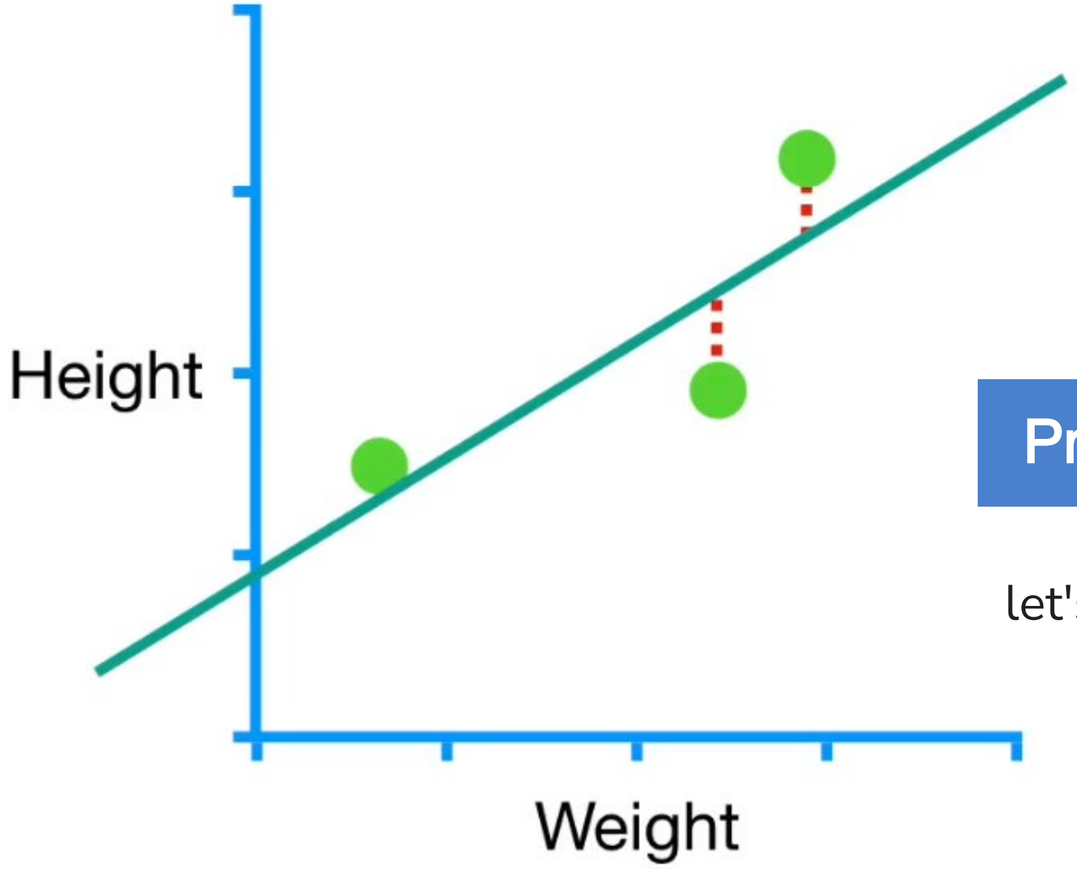
Step Size = **Slope** × **Learning Rate**

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

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

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





Now that we understand how Gradient Descent can estimate the Intercept

Predicted Height = intercept + slope × Weight

let's talk about how to estimate the Intercept and the Slope

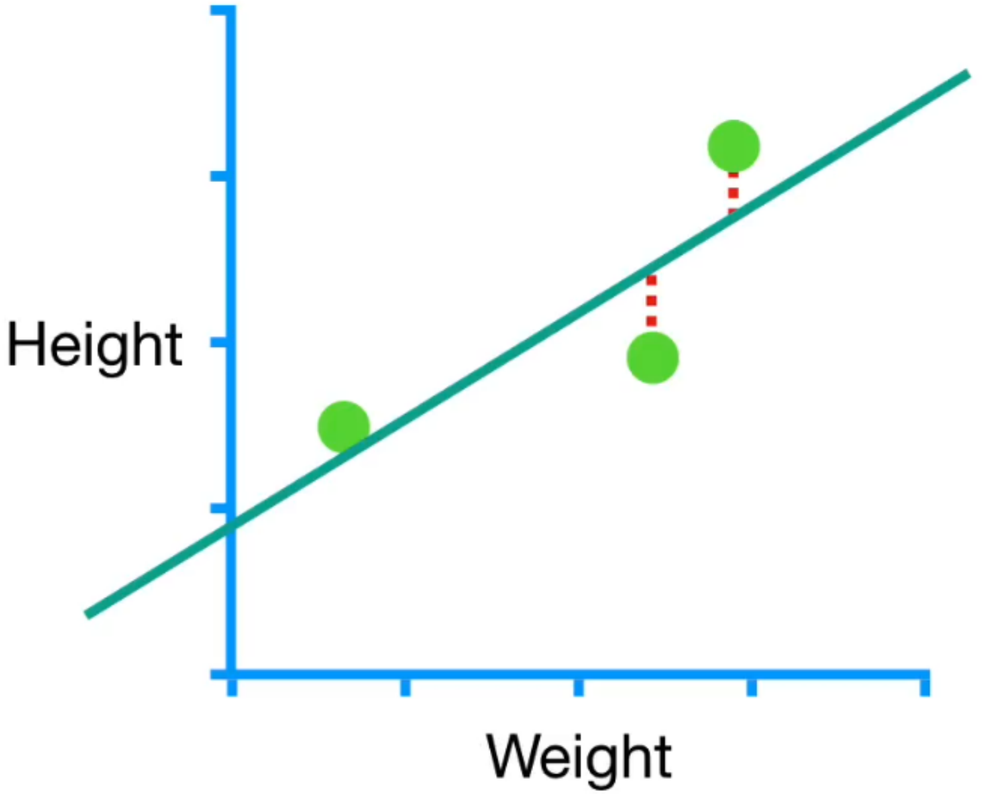
Just like before, we will use the Sum of the Squared Residuals as the **Loss Function**

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

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

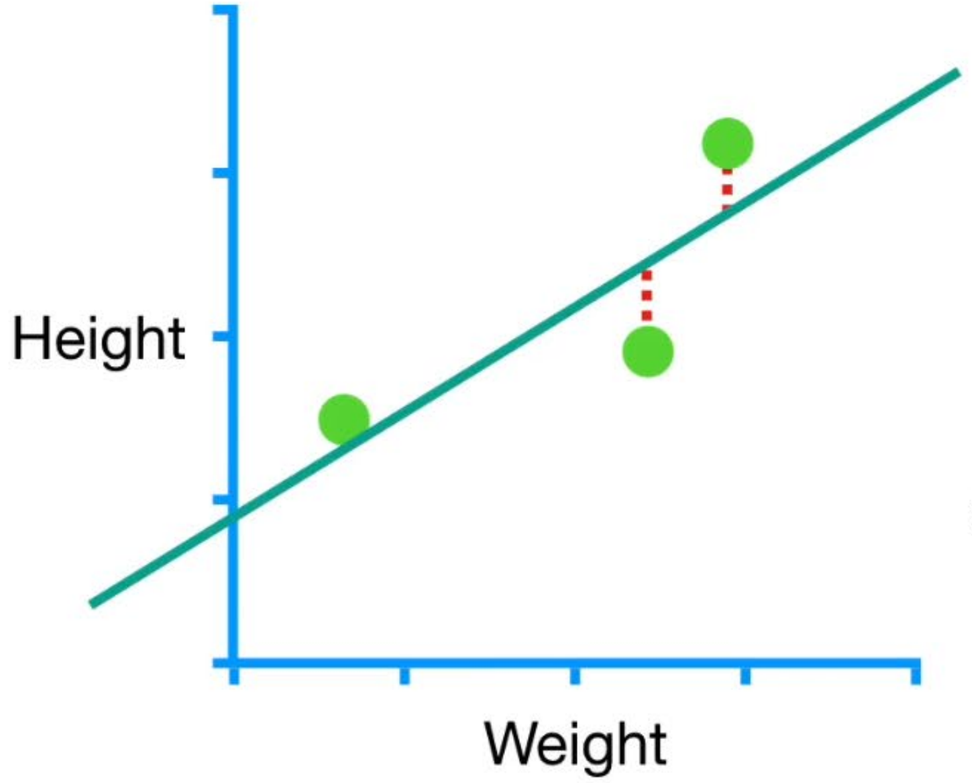


So, just like before, we need to take the derivative of this function…

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



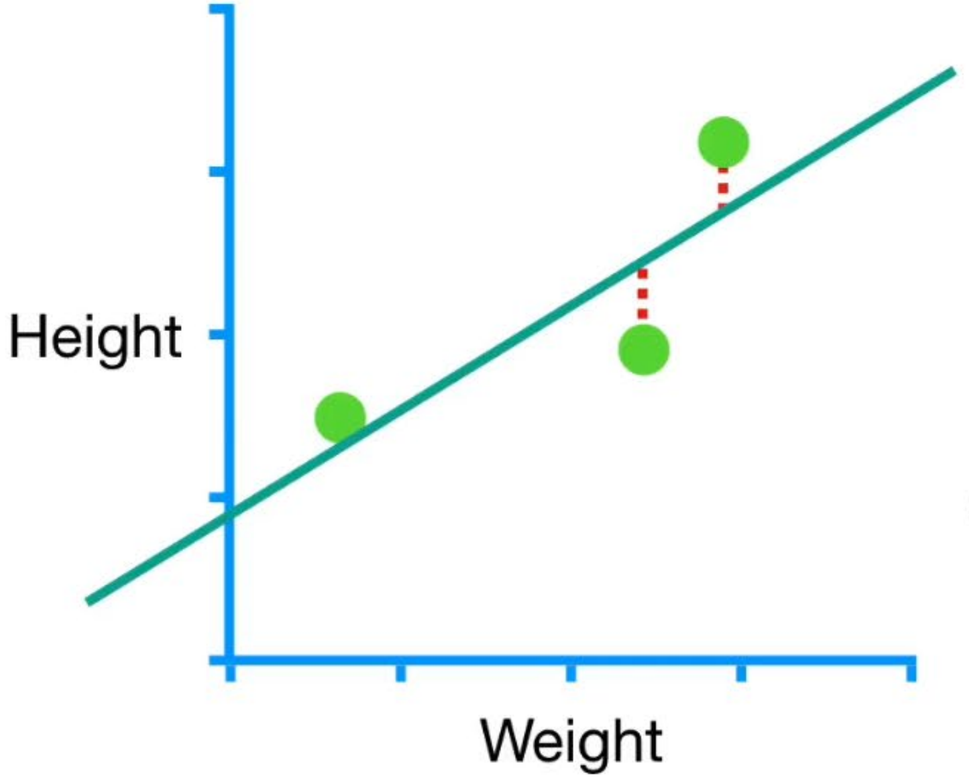
…and just like before, we'll take the derivative with respect to the **Intercept**…

$$\frac{d}{d\ intercept}$$ 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



…but unlike before, we'll also take the derivative with respect to the **Slope**!

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

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

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

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

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

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

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…

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

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

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

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

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

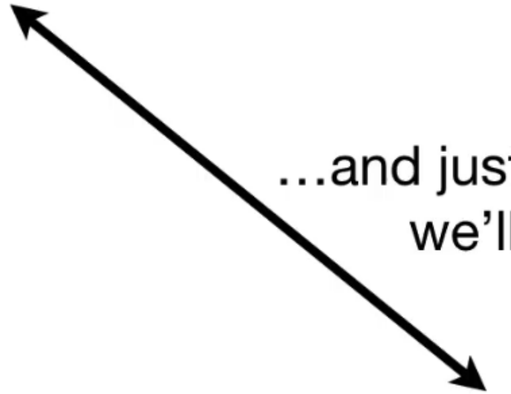
Just like before, we take the derivative of each part…

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

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

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

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



…and just like before,
we'll use…

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

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

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

The Chain Rule

The result is

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

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

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

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

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

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

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

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

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 take the derivative of each part…

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

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

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

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…

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

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

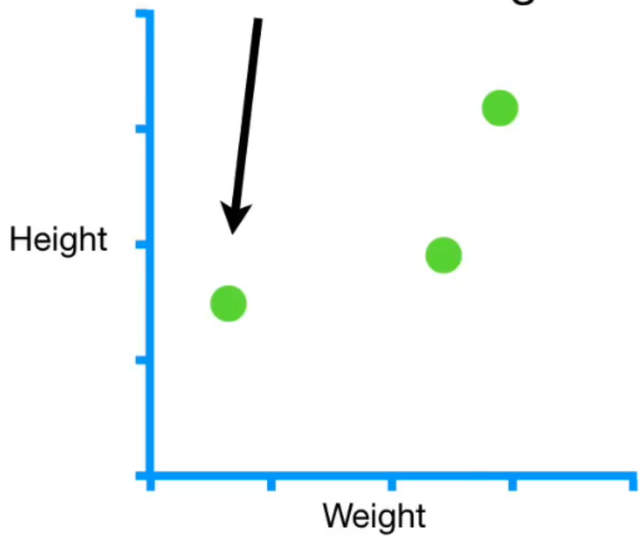
$$+ \frac{d}{d\,slope}(\mathbf{3.2} - (\text{intercept} + \text{slope} \times \mathbf{2.9}))^2$$

...and just like before, we'll use The Chain Rule

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

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

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.



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

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

…and this…

…is the derivative of the first part…

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

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

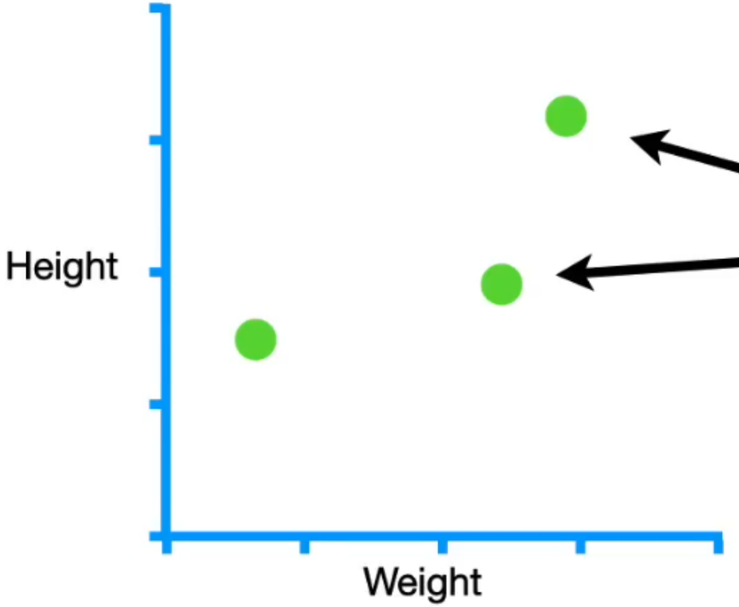
$$+ \frac{d}{d\ slope}(\mathbf{3.2} - (intercept + slope \times \mathbf{2.9}))^2$$

Likewise, we replace these terms with their derivatives.

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

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

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



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 =

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

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

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

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

...and here's the derivative with respect to the **Slope**.

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

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

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

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

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

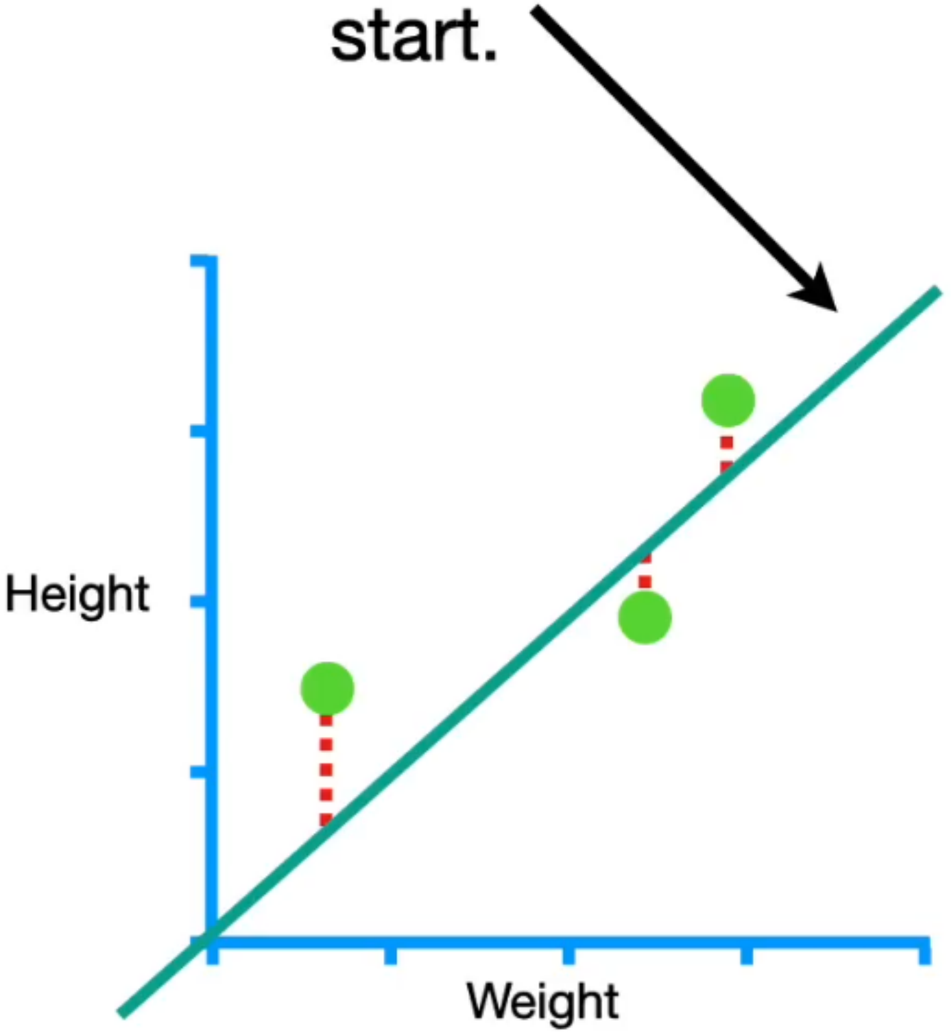
We will use this **Gradient** to **descend** to the lowest point in the **Loss Function**, which, in this case, is the Sum of the Squared Residuals.

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

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**.

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



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

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

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

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

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

$$\frac{d}{d\ slope}$$ 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 =

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

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

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

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

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

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

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

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

$$\frac{d}{d\,intercept}$$ 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 =

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Step Size$_{\text{Intercept}}$ = -1.6 × **Learning Rate**

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

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

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

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

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

Step Size$_{\text{Slope}}$ = -0.8 × **Learning Rate**

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

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

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

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

Step Size$_{\text{Intercept}}$ = -1.6 × 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.

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

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

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

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

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

This means that **Gradient Descent** can be very sensitive to the **Learning Rate**.

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

So, in theory, you shouldn't have to worry too much about the **Learning Rate**.

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

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

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

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

Step Size$_{Intercept}$ = -1.6 × 0.01 = **-0.016**

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

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

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

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

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

Step Size$_{Slope}$ = -0.8 × 0.01 = **-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 = Old Intercept - 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 = Old Slope - Step Size

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

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

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

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

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

New Intercept = 0 - **Step Size**

…and the
Step Sizes…

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

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

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

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

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

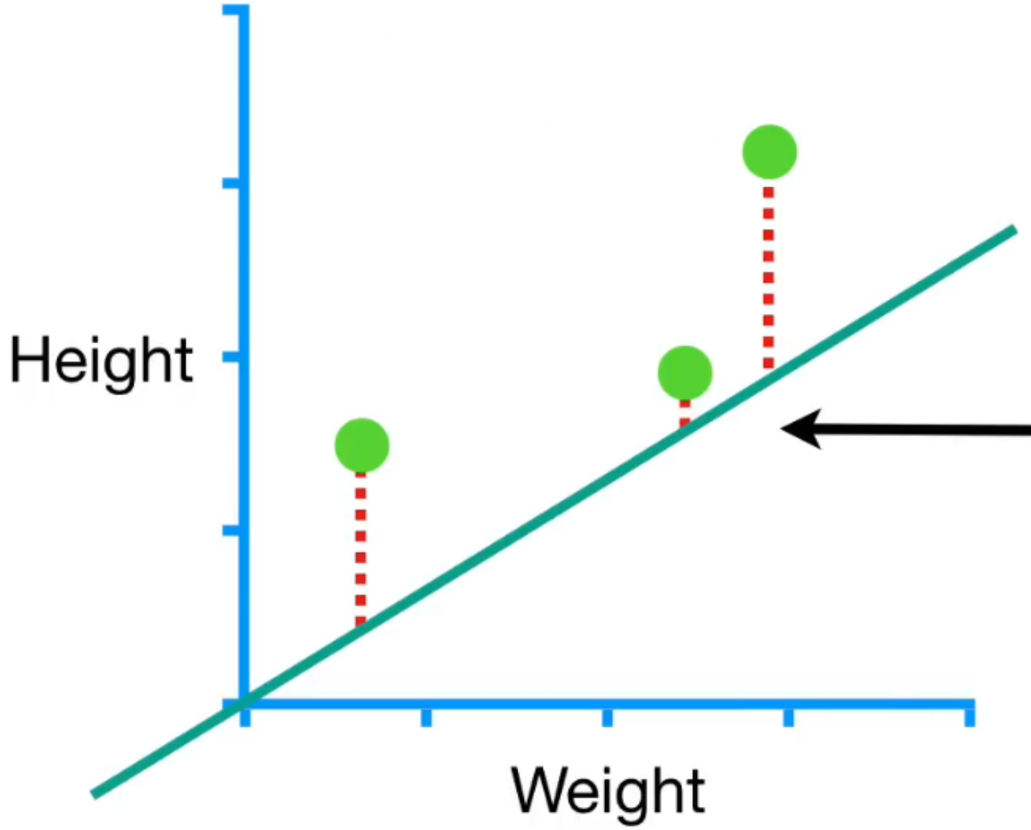
$$-2 \times \mathbf{0.5}(\mathbf{1.4} - (0 + 1 \times \mathbf{0.5}))$$

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

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

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

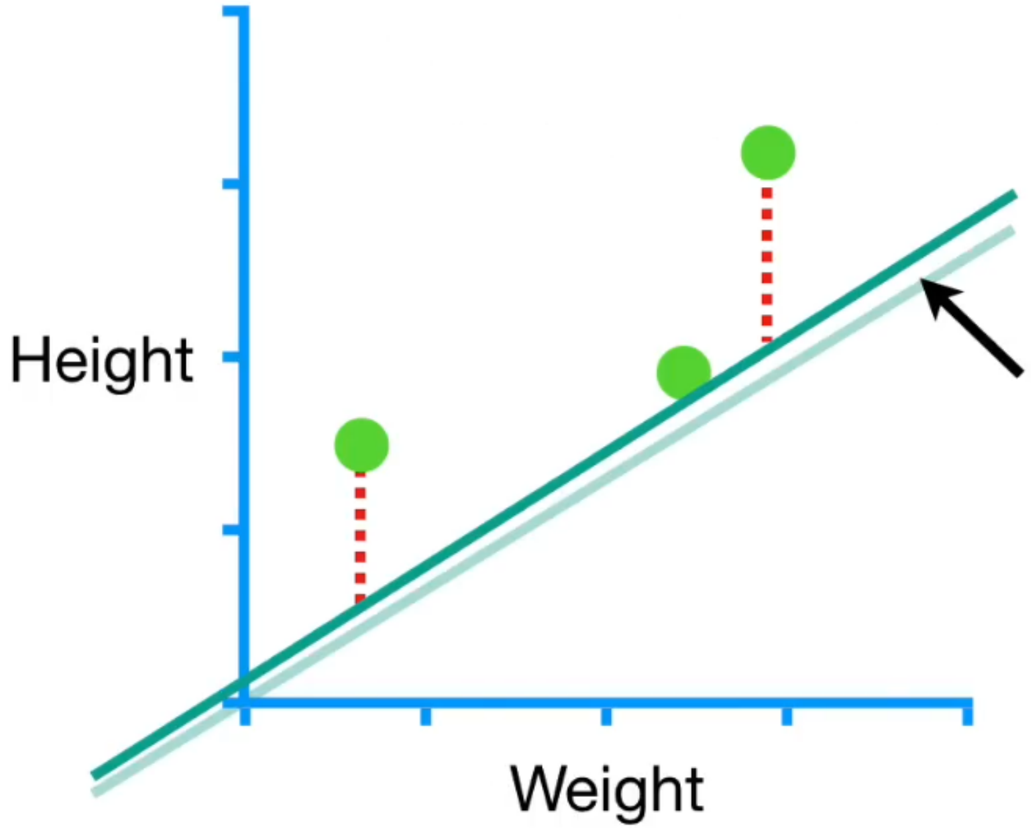
New Slope = 1 - (-0.008) = 1.008



New Intercept = 0 - (-0.016) = 0.016

This is the line we started with… (**Slope = 1** and **Intercept = 0**)

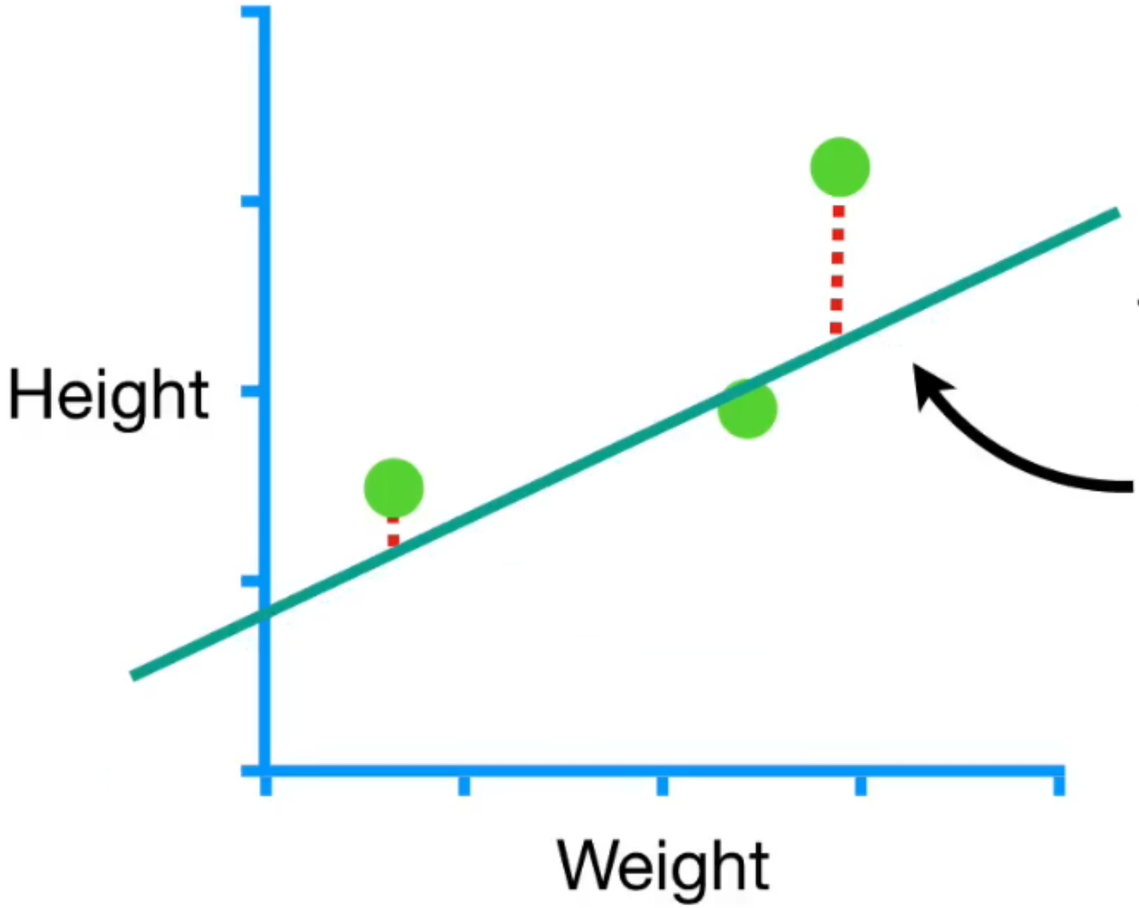
New Slope = 1 - (-0.008) = 1.008



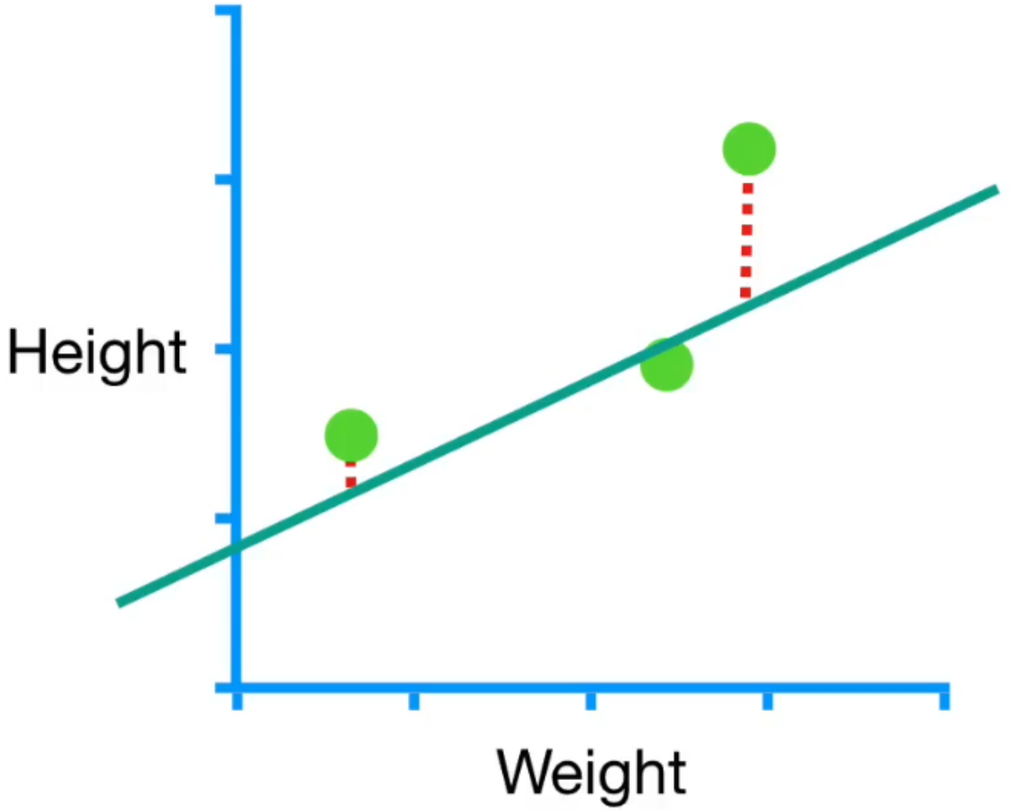
New Intercept = 0 - (-0.016) = 0.016

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

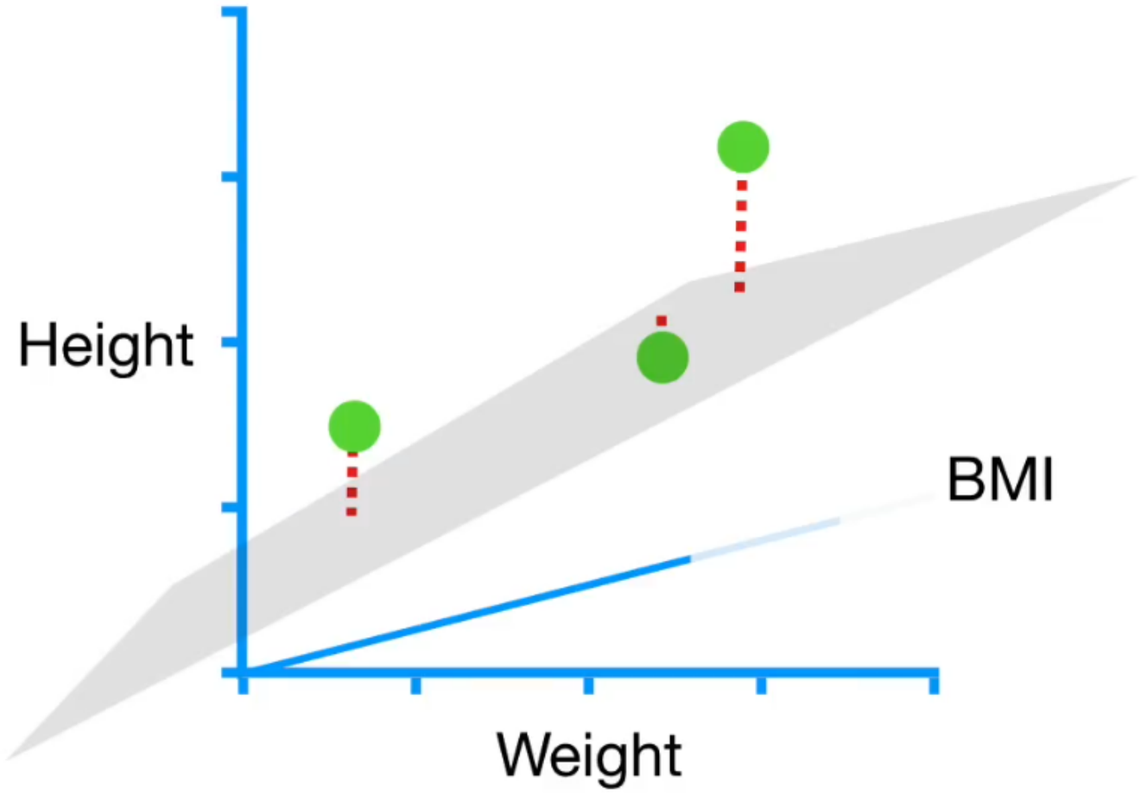
New Slope = 1 - (-0.008) = 1.008



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



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



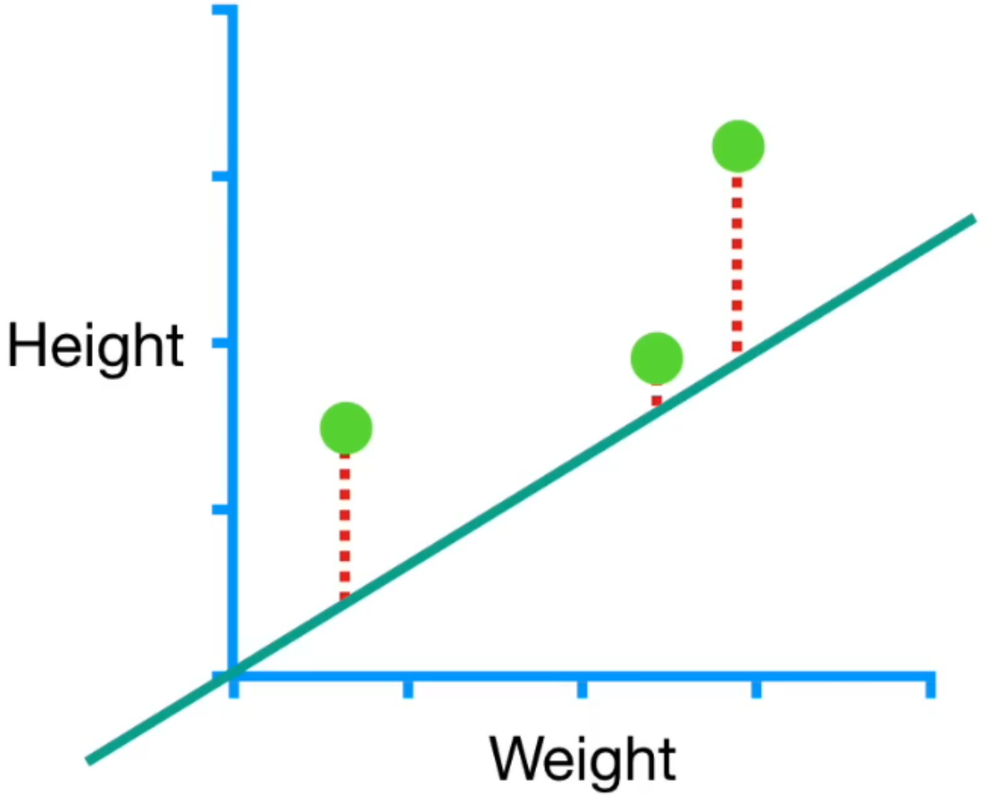
If we had more parameters, then we'd just take more derivatives and everything else stays the same.

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

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

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

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



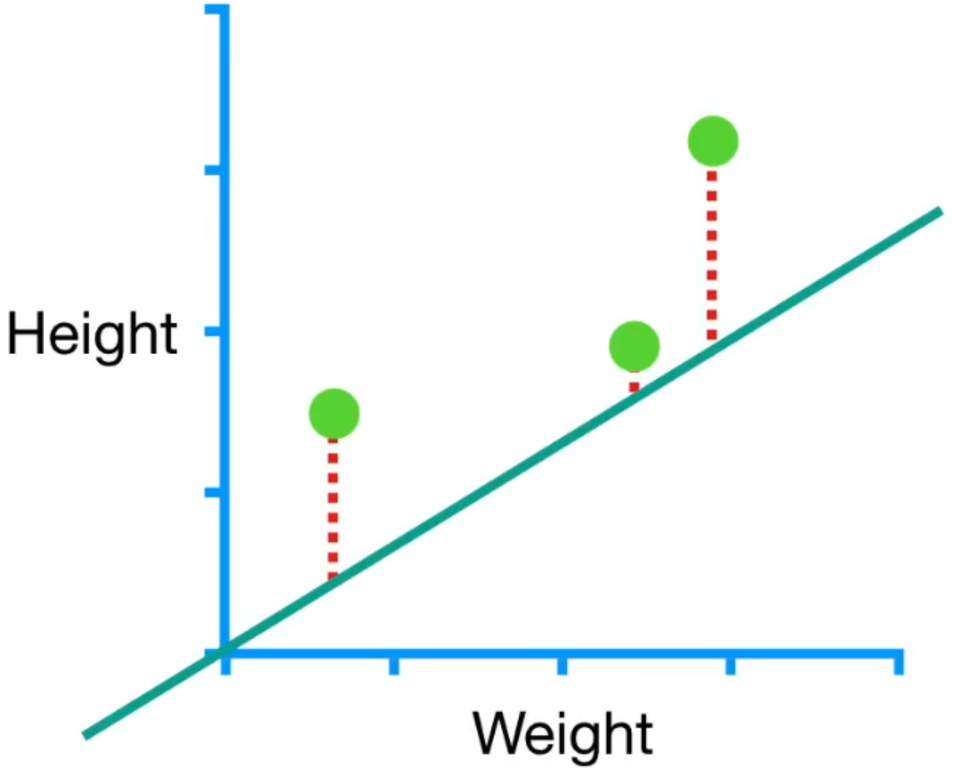
Sum of squared residuals = (**1.4** - (intercept + 0.64 x **0.5**))^2

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

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

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

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



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

Step 2: Pick random values for the parameters.

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

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

Step 5: Calculate the New Parameters:

New Parameter = Old Parameter - Step Size

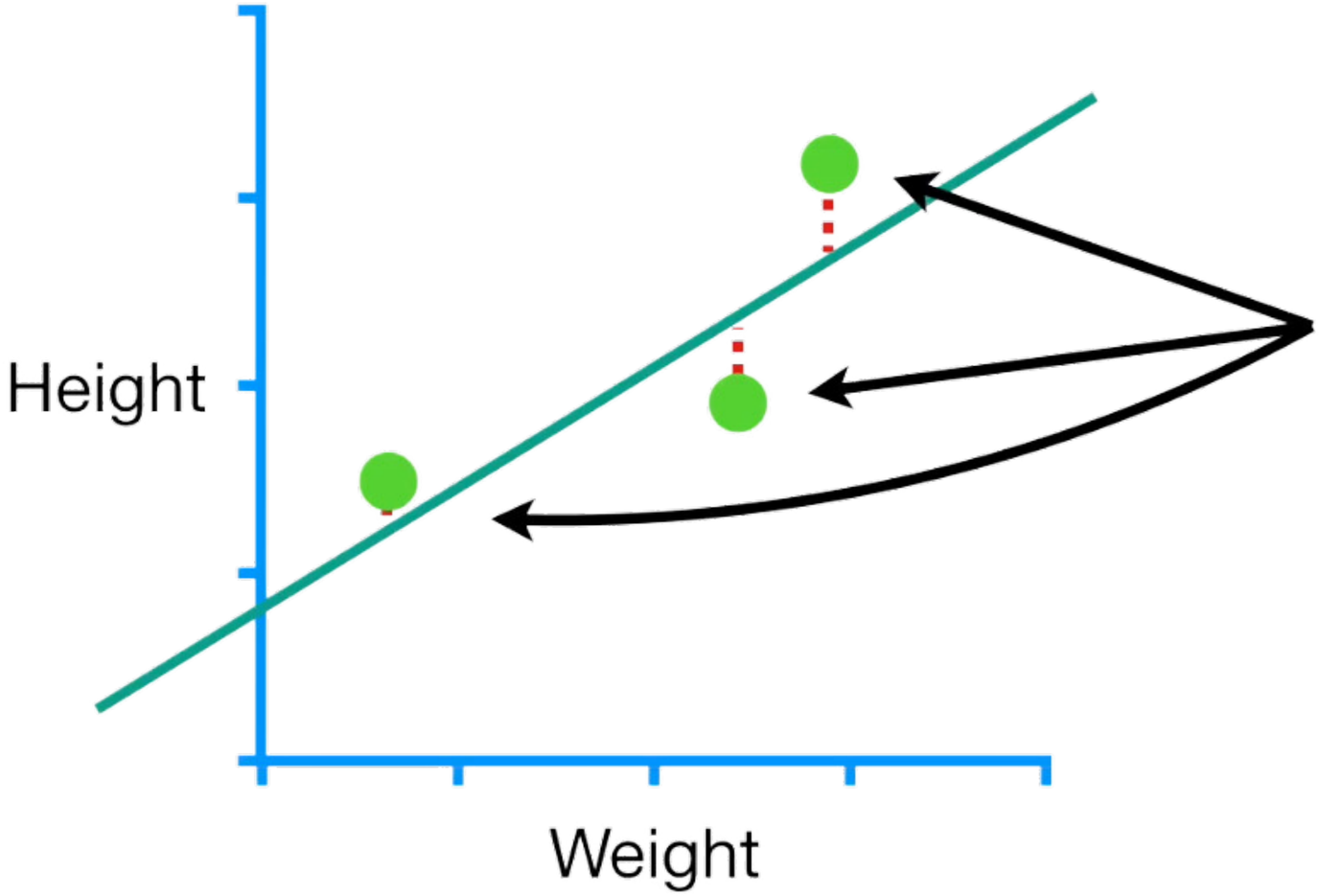
Now go back to Step 3 and repeat until Step Size is very small, or you reach the Maximum Number of Steps.

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 but when you have millions of data points, it can take a long time.

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.

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.

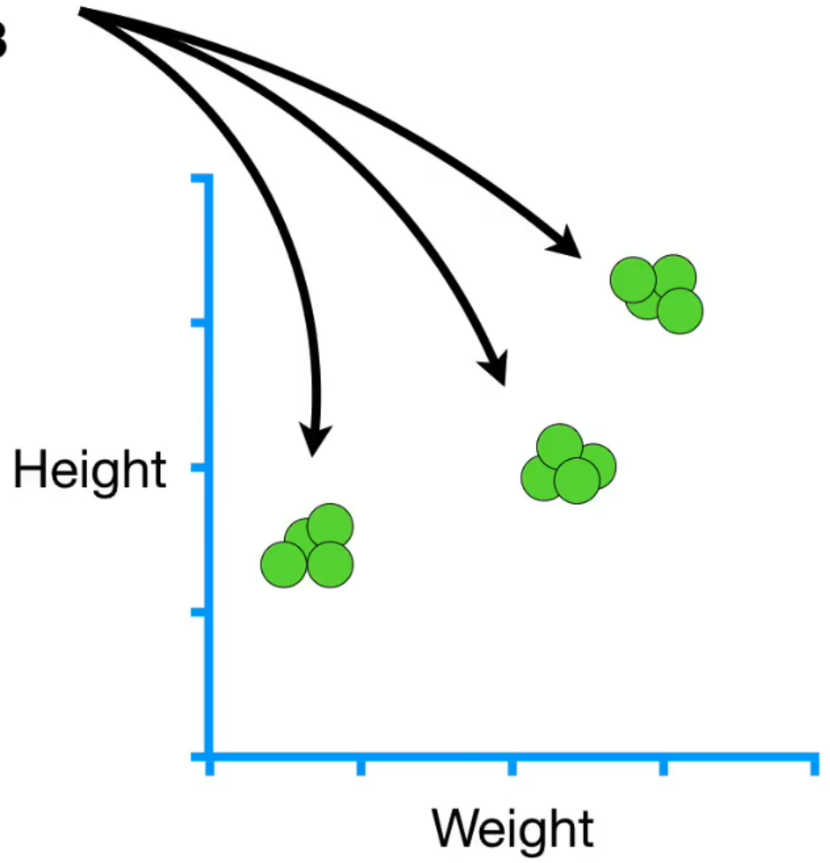
This reduces the time spent calculating the derivatives of the Loss Function.

Stochastic Gradient Descent sounds fancy, but it’s no big deal.

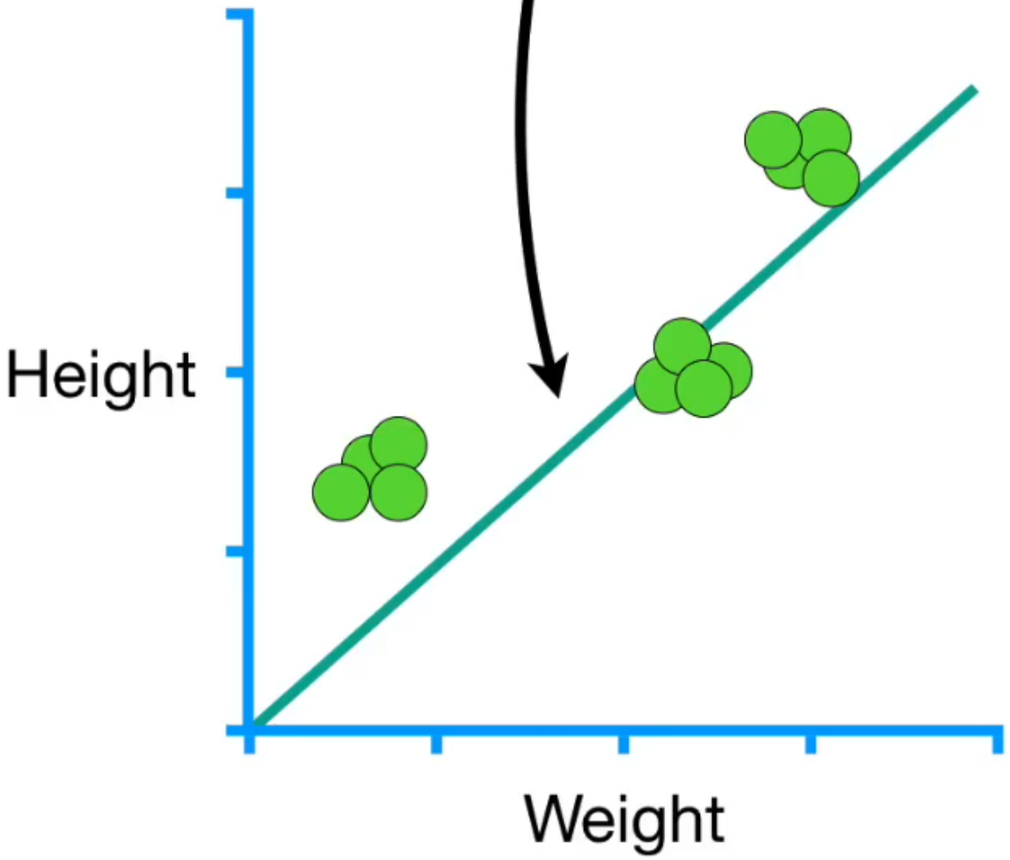
STOCHASTIC GRADIENT DESCENT

Stochastic Gradient Descent is especially useful when there are redundancies in the data.

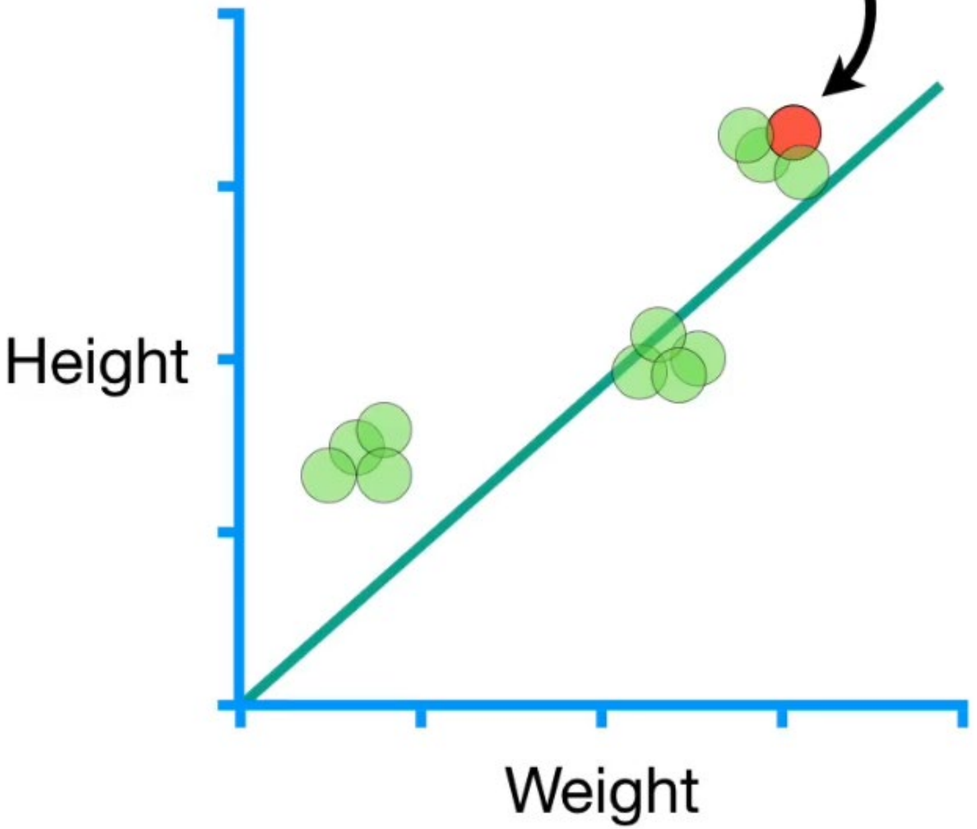
For example, we have **12** data points, but there is a lot of redundancy that forms **3** clusters.



So we start with a line with the **intercept** = **0** and the **slope** = **1**…



…then we randomly pick this point…



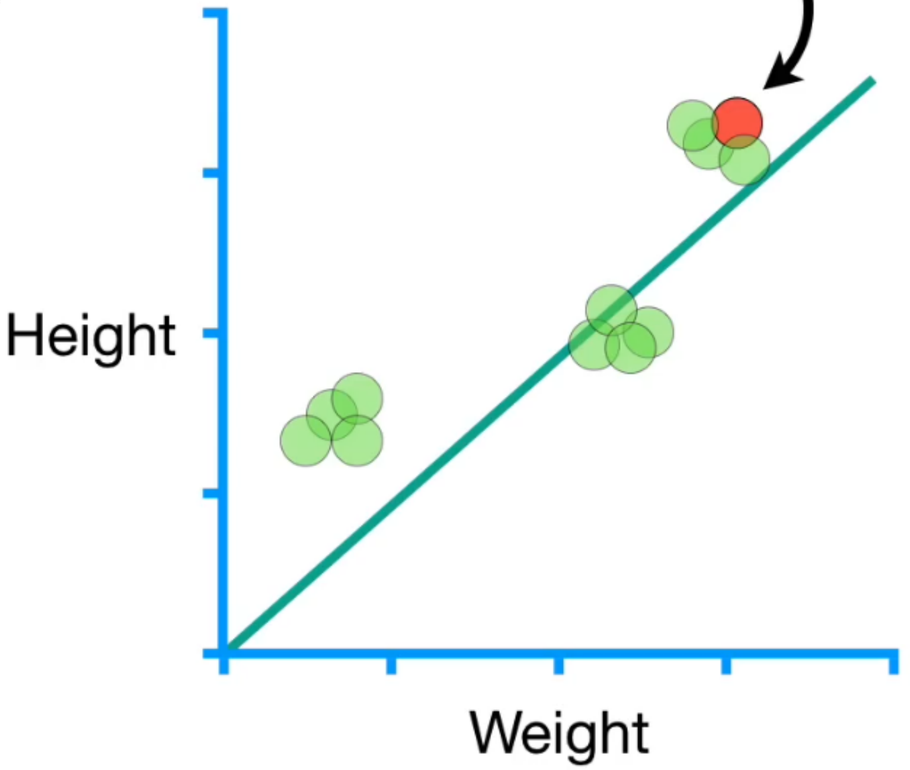
$$\frac{d}{d\ intercept}$$ Sum of squared residuals =

-2(**Height** - (0 + 1 × **Weight**)

…so we plug in the **Weight**, **3**…

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

-2 × **Weight**(**Height** - (0 + 1 × **Weight**))



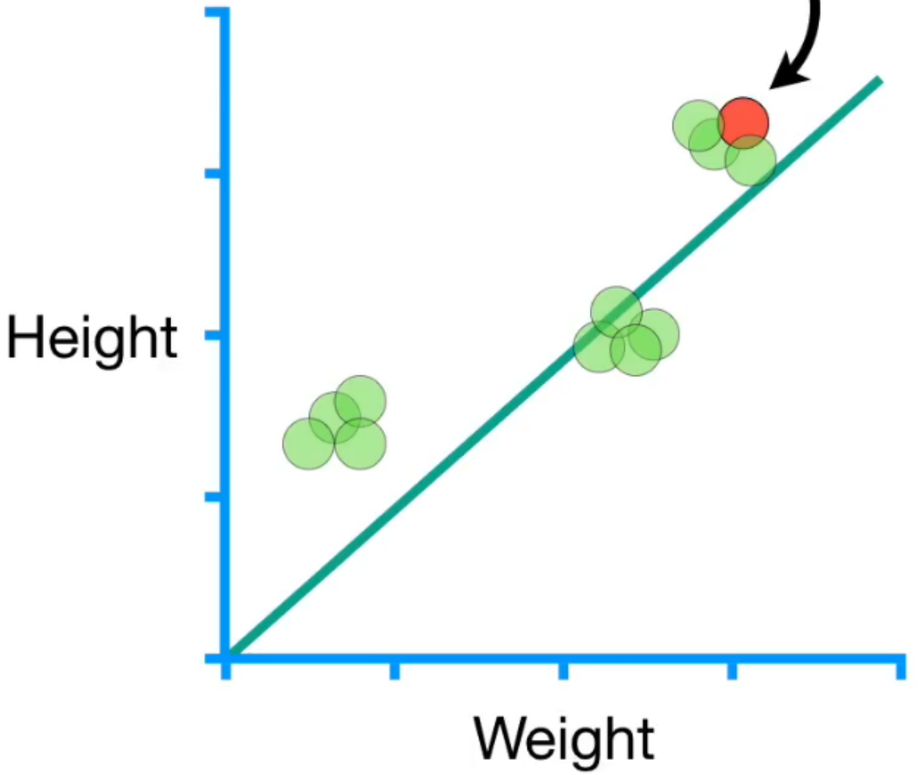
$\frac{d}{d\ intercept}$ Sum of squared residuals =

-2(**Height** - (0 + 1 × **3**)

…so we plug in the **Weight**, **3**…

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

-2 × **3**(**Height** - (0 + 1 × **3**))



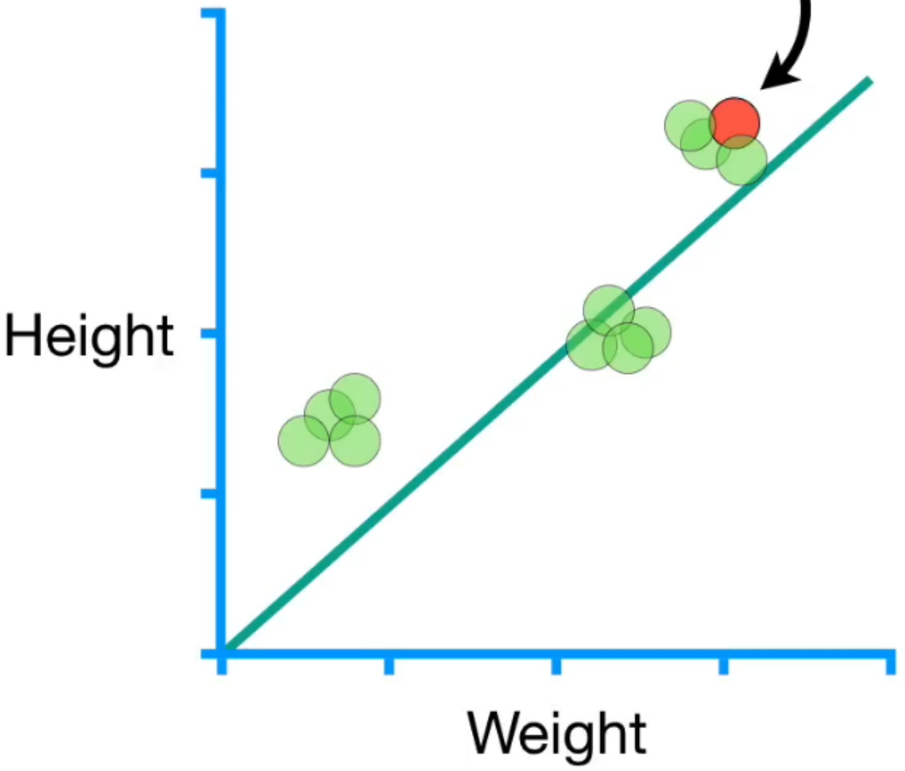
$\frac{d}{d\ intercept}$ Sum of squared residuals =

$-2(\mathbf{3.3} - (0 + 1 \times \mathbf{3})$

…and **Height**, **3.3**…

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

$-2 \times \mathbf{3}(\mathbf{3.3} - (0 + 1 \times \mathbf{3}))$



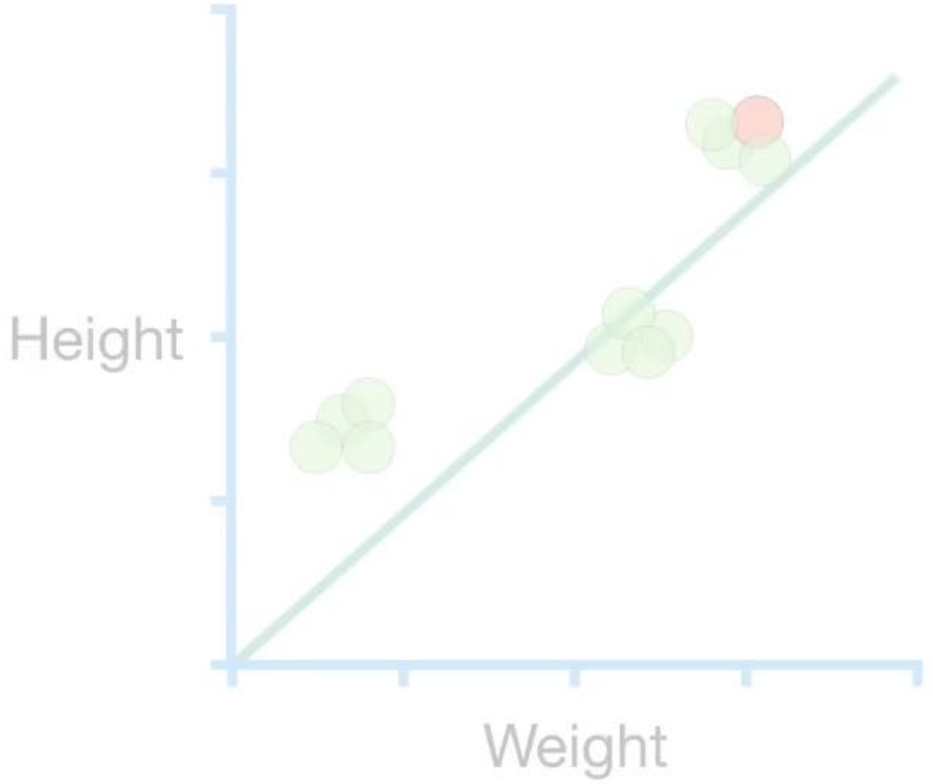
$$\frac{d}{d\,intercept}\text{Sum of squared residuals} = -2(\mathbf{3.3} - (0 + 1 \times \mathbf{3}) = -0.6$$

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

…plug in the slopes…

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

$$\frac{d}{d\,slope}\text{Sum of squared residuals} = -2 \times \mathbf{3}(\mathbf{3.3} - (0 + 1 \times \mathbf{3})) = -1.8$$



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

$-2(\mathbf{3.3} - (0 + 1 \times \mathbf{3}) = -0.6$

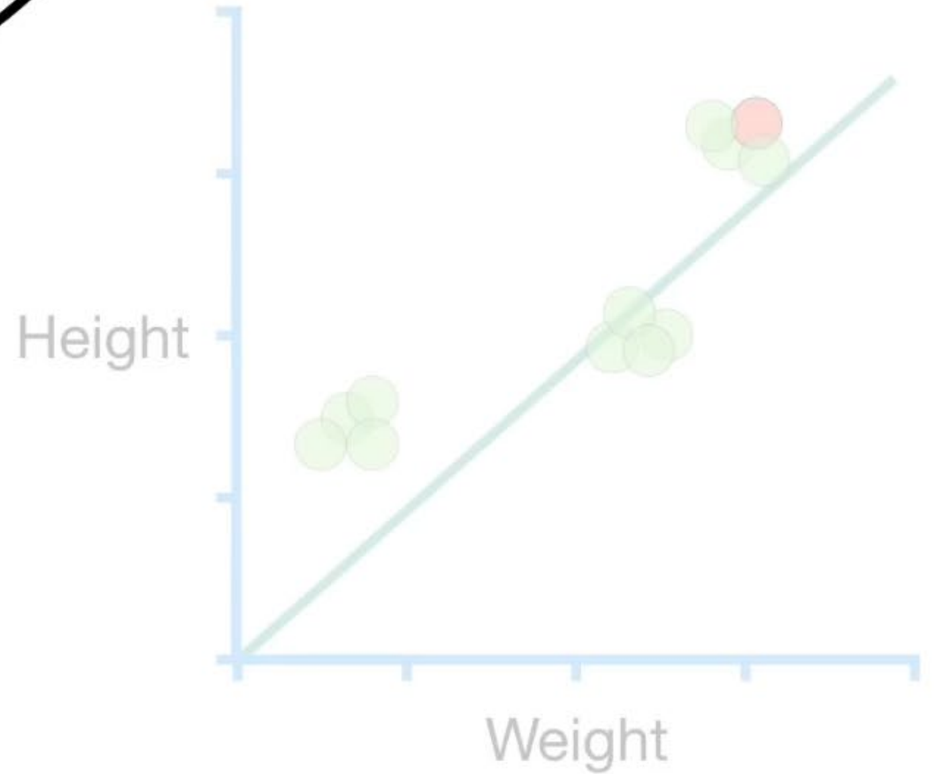
…then multiply by the **Learning Rate**.

Step Size$_{\text{Intercept}}$ = -0.6 × **Learning Rate**

Step Size$_{\text{Slope}}$ = -1.8 × **Learning Rate**

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

$-2 \times \mathbf{3}(\mathbf{3.3} - (0 + 1 \times \mathbf{3})) = -1.8$



$$\frac{d}{d\,intercept}\text{Sum of squared residuals} = -2(\mathbf{3.3} - (0 + 1 \times \mathbf{3}) = -0.6$$

$$\textbf{Step Size}_{\text{Intercept}} = -0.6 \times \textbf{Learning Rate}$$

$$\textbf{Step Size}_{\text{Slope}} = -1.8 \times \textbf{Learning Rate}$$

$$\frac{d}{d\,slope}\text{Sum of squared residuals} = -2 \times \mathbf{3}(\mathbf{3.3} - (0 + 1 \times \mathbf{3})) = -1.8$$

NOTE: Just like with regular **Gradient Descent**, **Stochastic Gradient Descent** is sensitive to the value you choose for the **Learning Rate**…

…and just like for regular **Gradient Descent**, the general strategy is to start with a relatively *large* **Learning Rate** and make it *smaller* with each step…

…and lastly, just like for regular **Gradient Descent**, many implementations of **Stochastic Gradient Descent** will take care of this for you by default.

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

$-2(\mathbf{3.3} - (0 + 1 \times \mathbf{3}) = -0.6$

Step Size$_{Intercept}$ = -0.6 × **Learning Rate**

Step Size$_{Slope}$ = -1.8 × **Learning Rate**

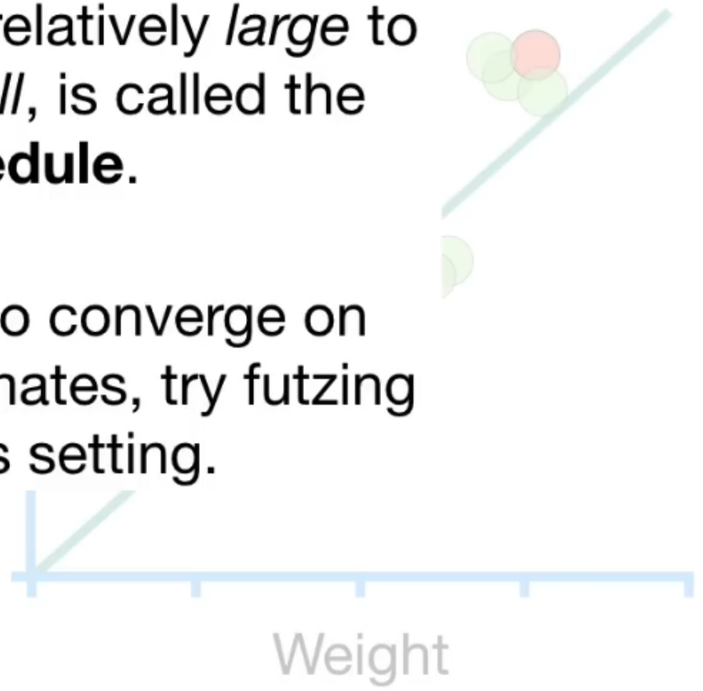
$\frac{d}{d\,slope}$ Sum of squared residuals =

$-2 \times \mathbf{3}(\mathbf{3.3} - (0 + 1 \times \mathbf{3})) = -1.8$

TERMINOLOGY ALERT!!!

The way the **Learning Rate** changes, from relatively *large* to relatively *small*, is called the **schedule**.

So if you fail to converge on parameter estimates, try futzing with this setting.



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

$$-2(\mathbf{3.3} - (0 + 1 \times \mathbf{3}) = -0.6$$

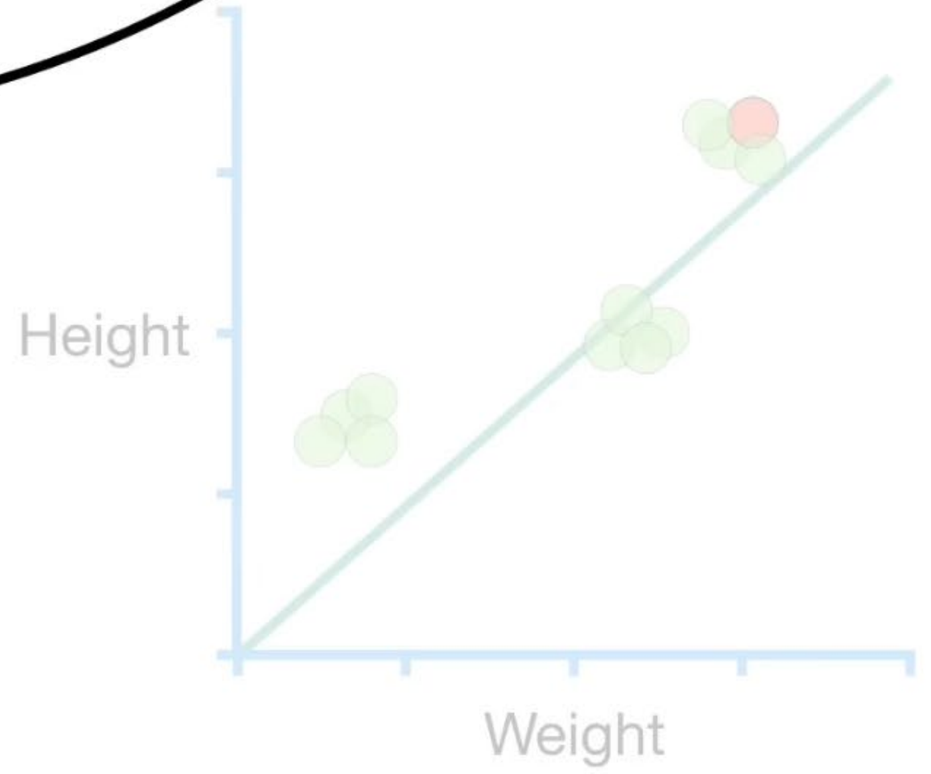
In this simple example, however, we'll just setting the **Learning Rate** to **0.01**.

Step Size$_{Intercept}$ = -0.6 × **Learning Rate**

Step Size$_{Slope}$ = -1.8 × **Learning Rate**

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

$$-2 \times \mathbf{3}(\mathbf{3.3} - (0 + 1 \times \mathbf{3})) = -1.8$$



The result is

$$\textbf{Step Size}_{\text{Intercept}} = -0.6 \times 0.01 = \mathbf{-0.006}$$

$$\textbf{Step Size}_{\text{Slope}} = -1.8 \times 0.01 = \mathbf{-0.018}$$

then, Calculate the new intercept

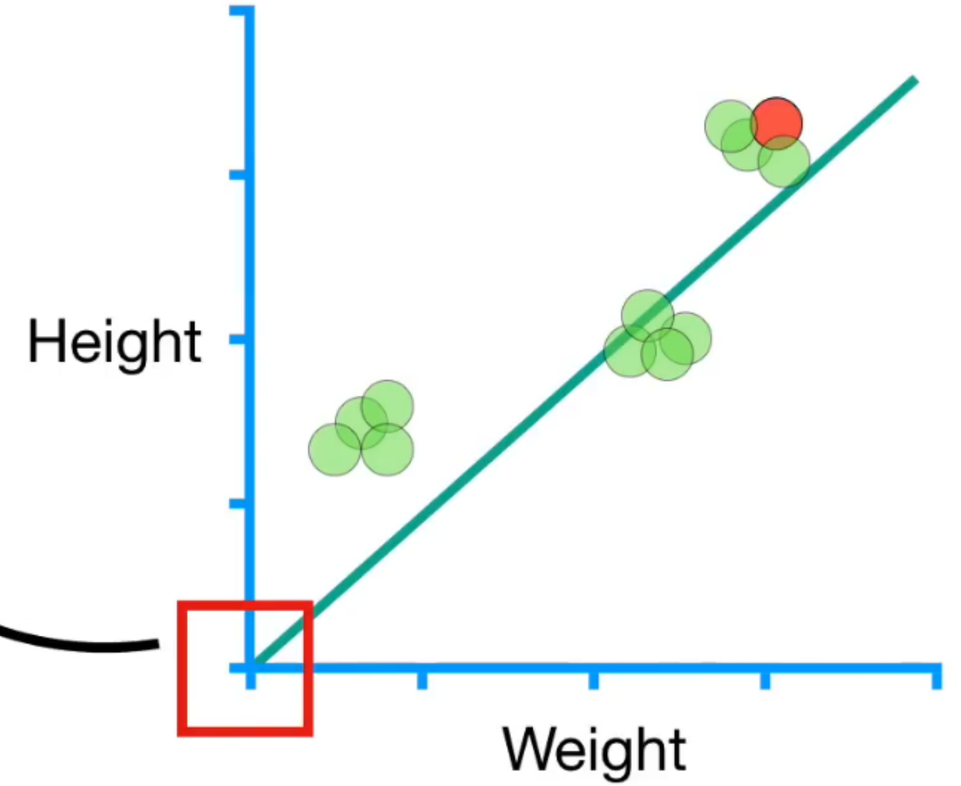
New Intercept = Old Intercept - Step Size

New Intercept = 0 - **Step Size**

Step Size$_{\text{Intercept}}$ = -0.6 x 0.01 = **-0.006**

Step Size$_{\text{Slope}}$ = -1.8 x 0.01 = **-0.018**

…calculate the new **intercept**…

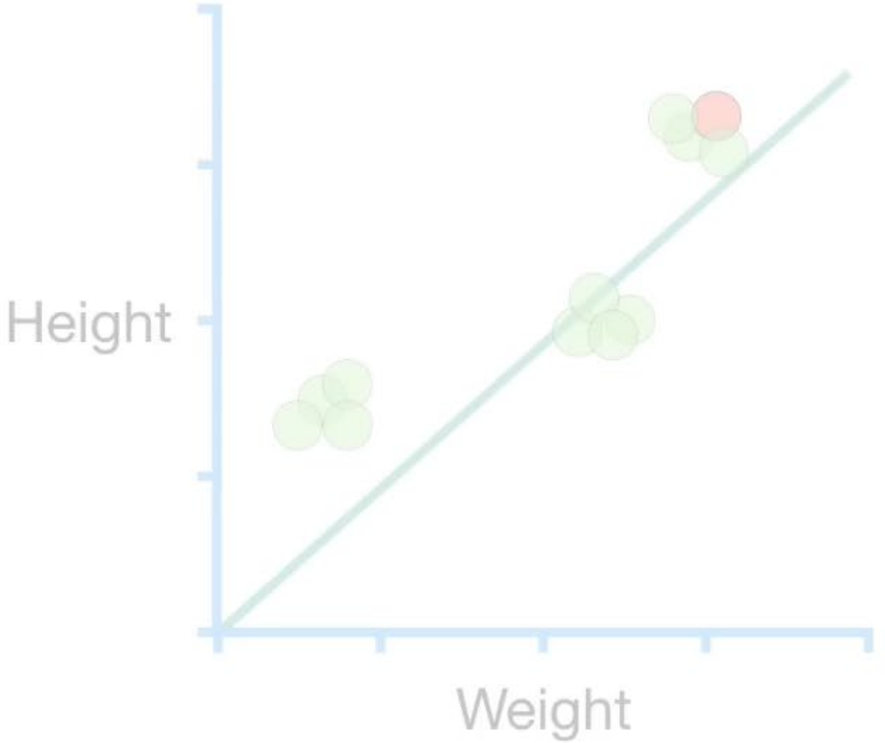


New Intercept = 0 - **Step Size**

Step Size$_{\text{Intercept}}$ = -0.6 x 0.01 = **-0.006**

Step Size$_{\text{Slope}}$ = -1.8 x 0.01 = **-0.018**

…calculate the new **intercept**…



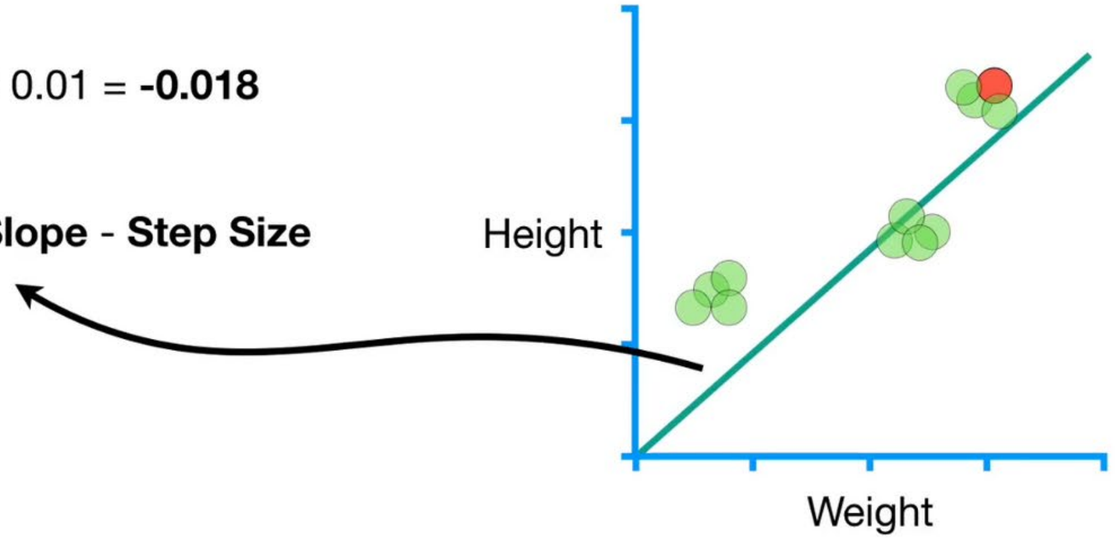
New Intercept = 0 - -0.006 = 0.006

Step Size$_{\text{Intercept}}$ = -0.6 x 0.01 = **-0.006**

Step Size$_{\text{Slope}}$ = -1.8 x 0.01 = **-0.018**

New Slope = **Old Slope** - **Step Size**

…and the new **slope**.



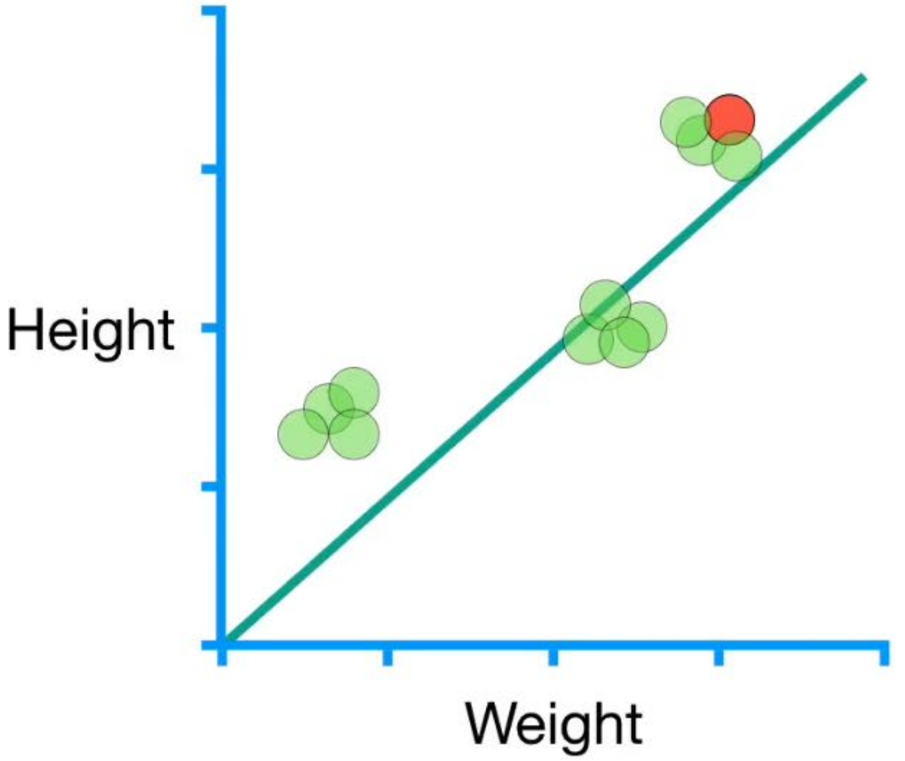
New Intercept = 0 - -0.006 = 0.006

Step Size$_{\text{Intercept}}$ = -0.6 x 0.01 = **-0.006**

Step Size$_{\text{Slope}}$ = -1.8 x 0.01 = **-0.018**

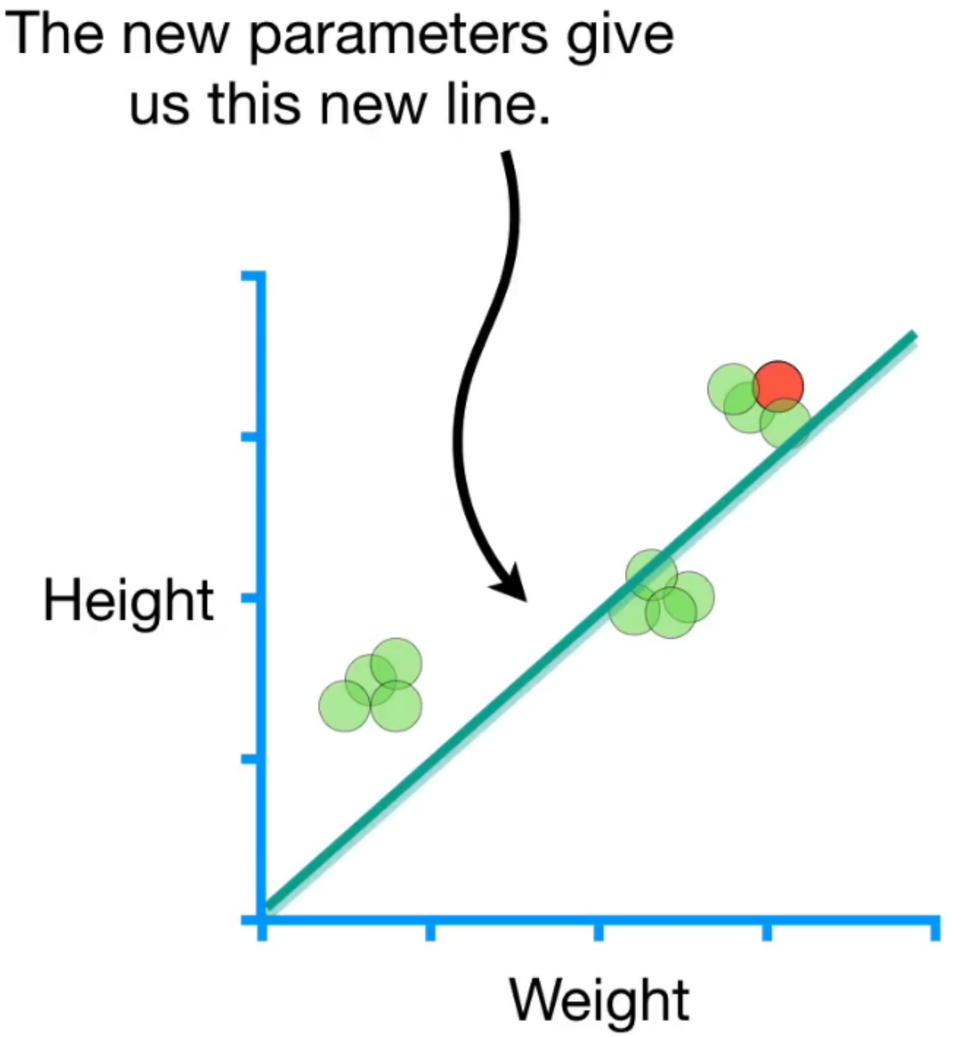
New Slope = 1 - **Step Size**

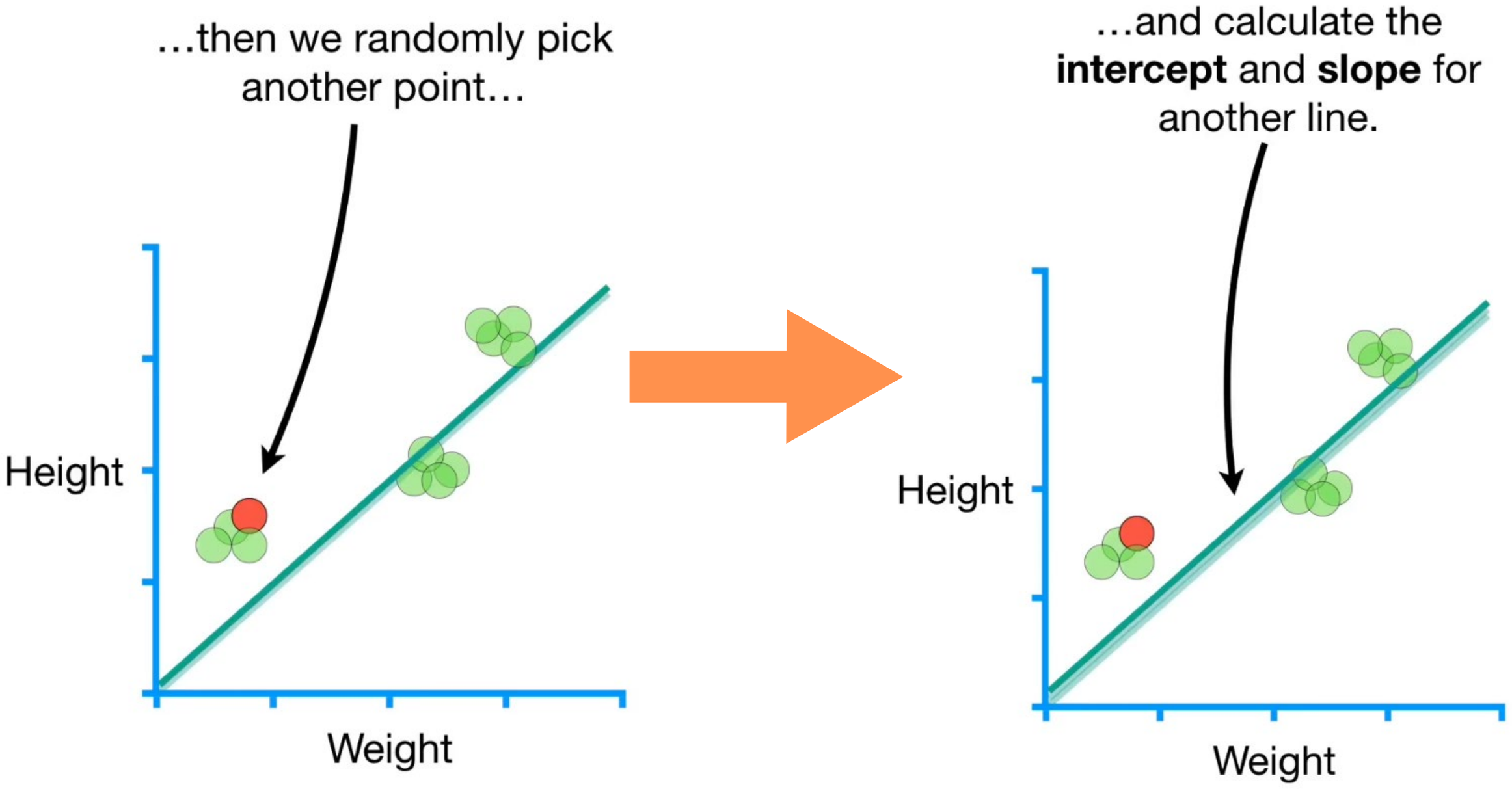
…and the new **slope**.



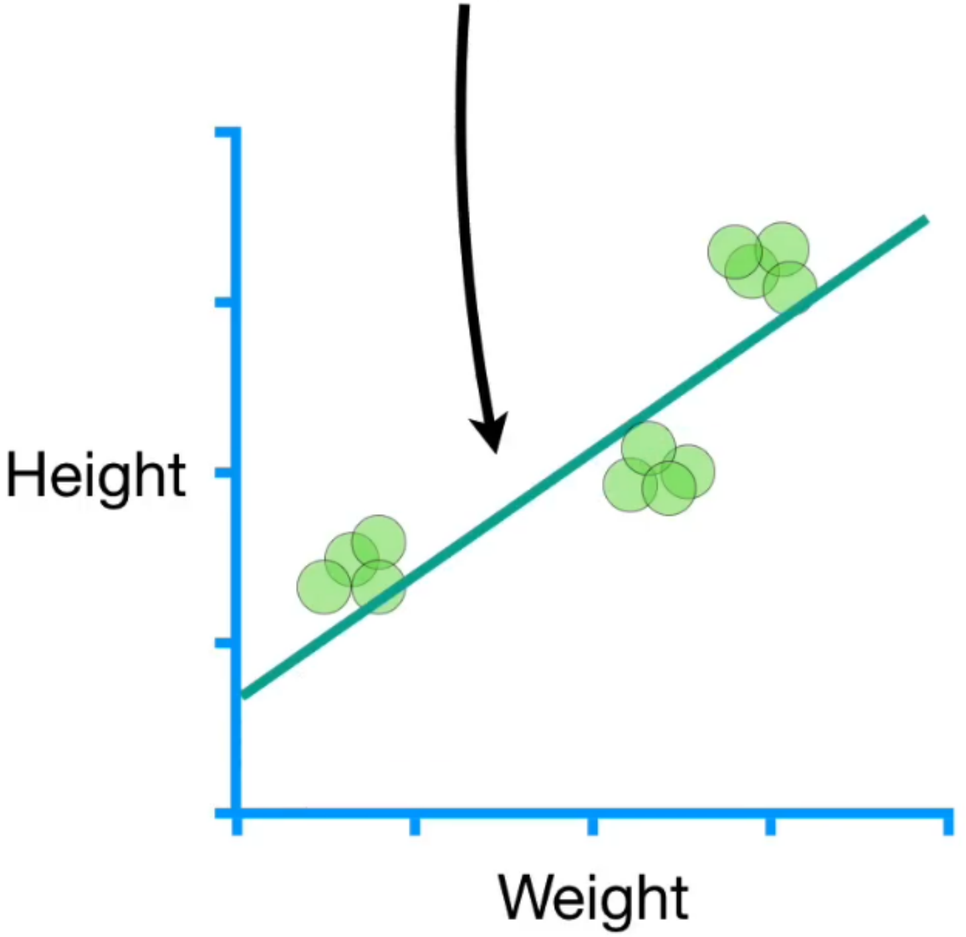
New Intercept = 0 - -0.006 = 0.006

New Slope = 1 - -0.018 = 1.018



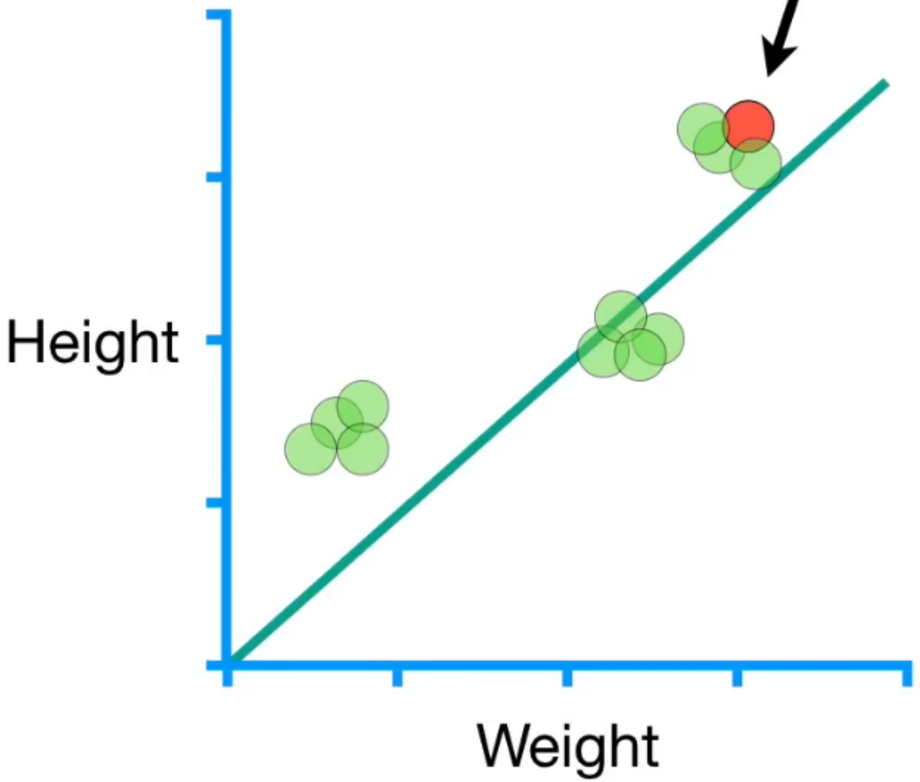

…then we randomly pick another point…
Height
Weight
…and calculate the **intercept** and **slope** for another line.
Height
Weight

…and ultimately we end up with a line where the **intercept** = **0.85** and the **slope** = **0.68**.

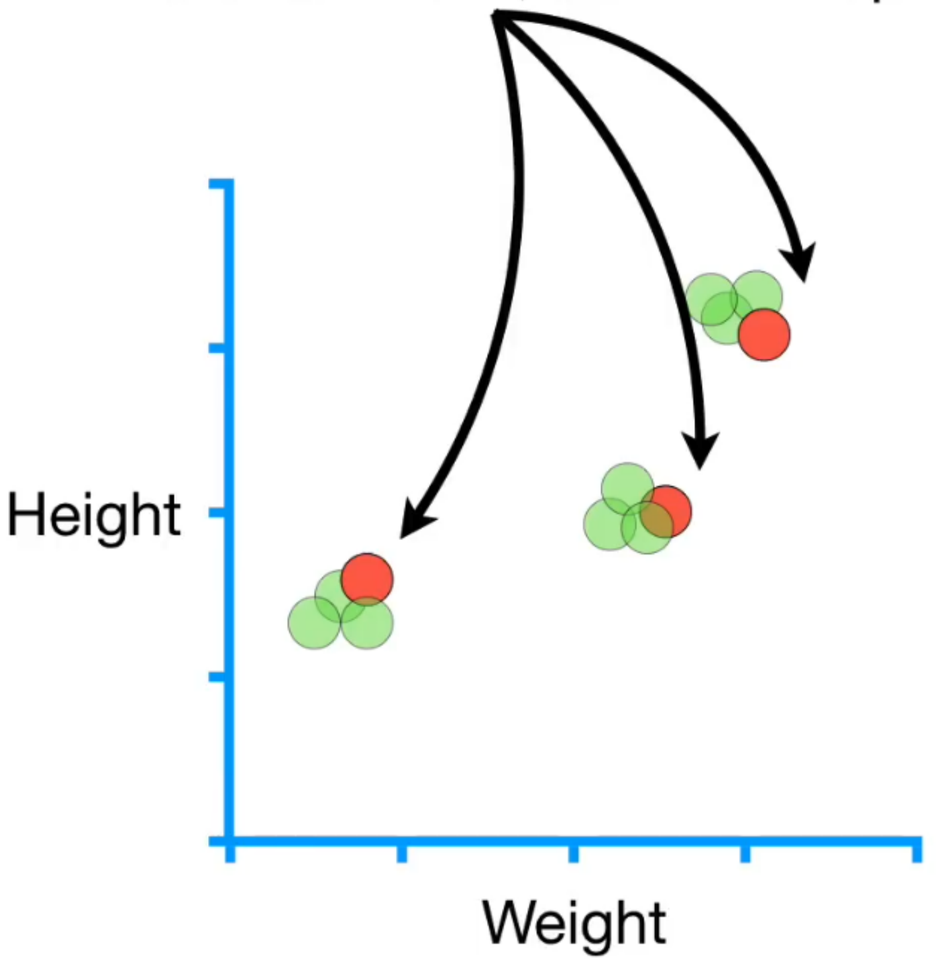


…and the least squares estimates, aka, the gold standard, gives a line where the **intercept** = **0.87** and the **slope** = **0.68**.

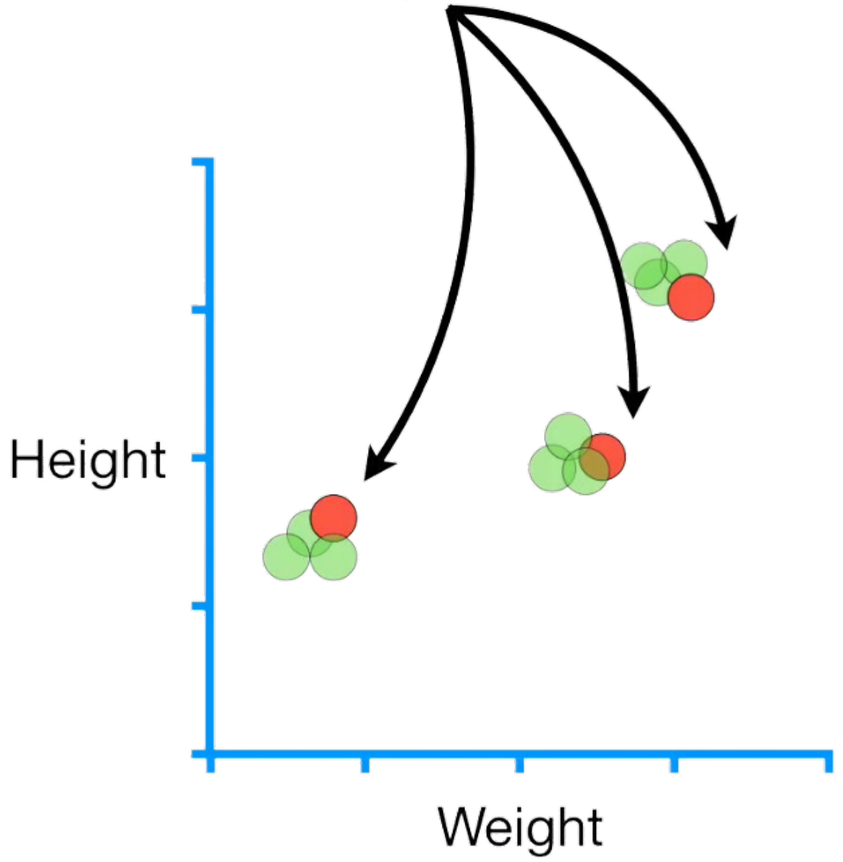
NOTE: The strict definition of **Stochastic Gradient Descent** is to only use **1** sample per step…



…however, it is more common to select a small subset of data, or **mini-batch**, for each step.



For example, we could use **3** samples per step, instead of just **1**.

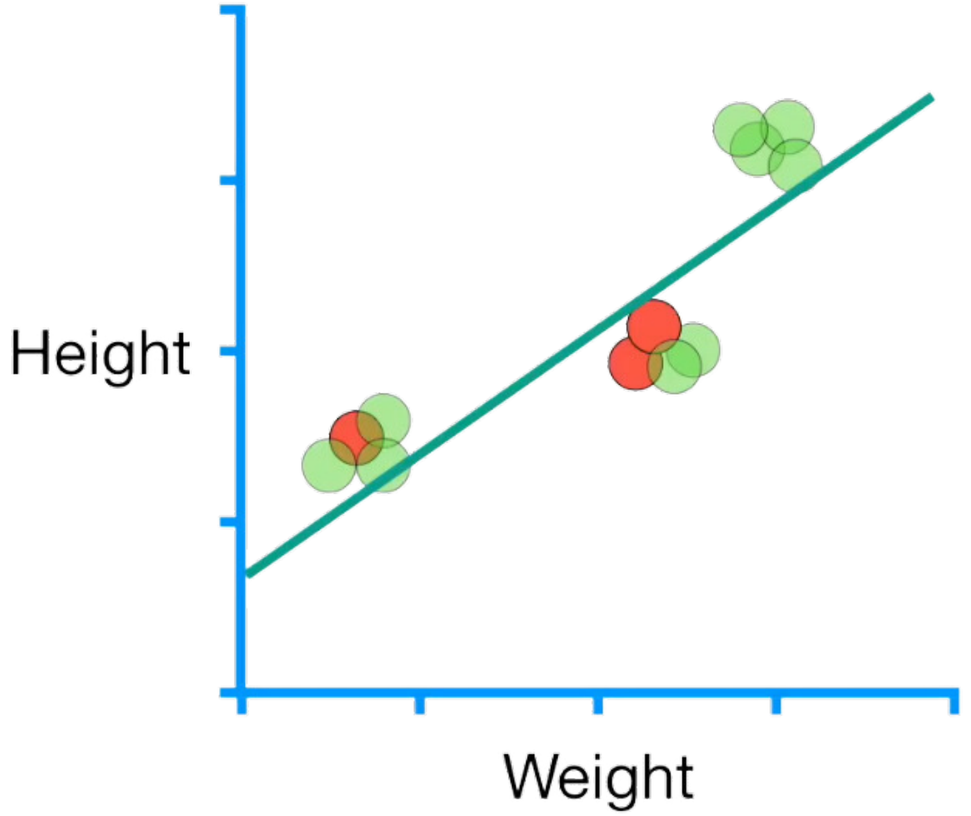


Using a **mini-batch** for each step takes the best of both worlds between using just one sample and all of of the data at each step.

Similar to using all of the data, using a **mini-batch** can result in more stable estimates of the parameters in fewer steps…

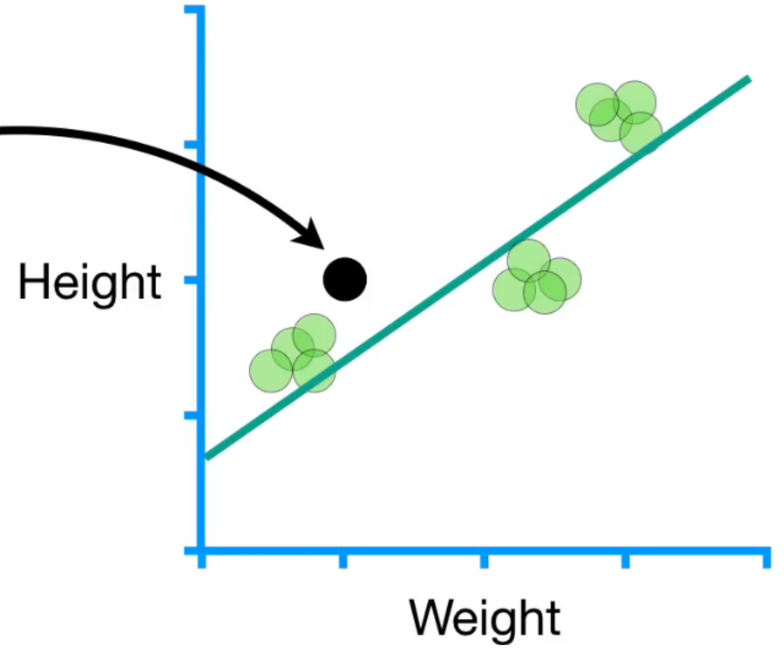
…and like using just one sample per step, using a **mini-batch** is much faster than using all of the data.

In this example, using **3** samples per step we ended up with the **intercept** = **0.86** and the **slope** = **0.68**.

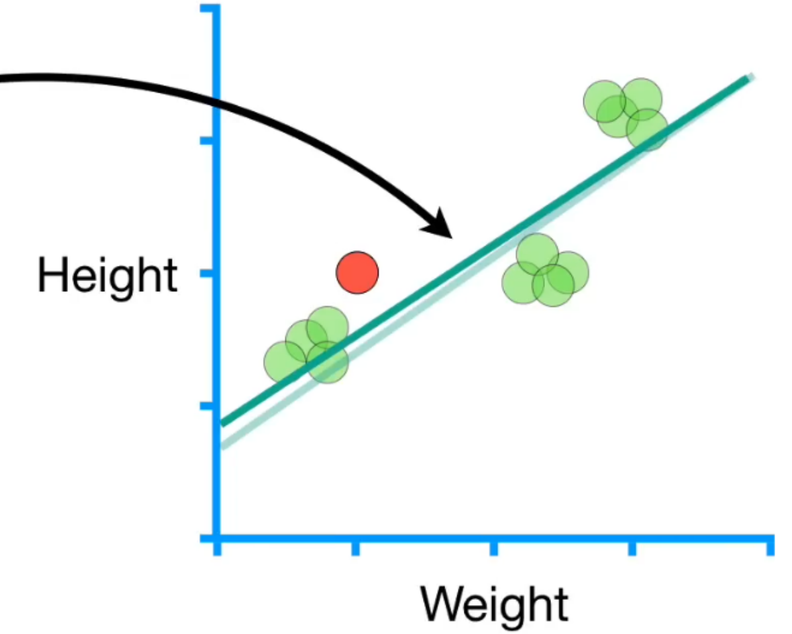


…which means that the estimate for the intercept was just a little closer to the gold standard, **0.87**, then when we used one sample and got **0.85**.

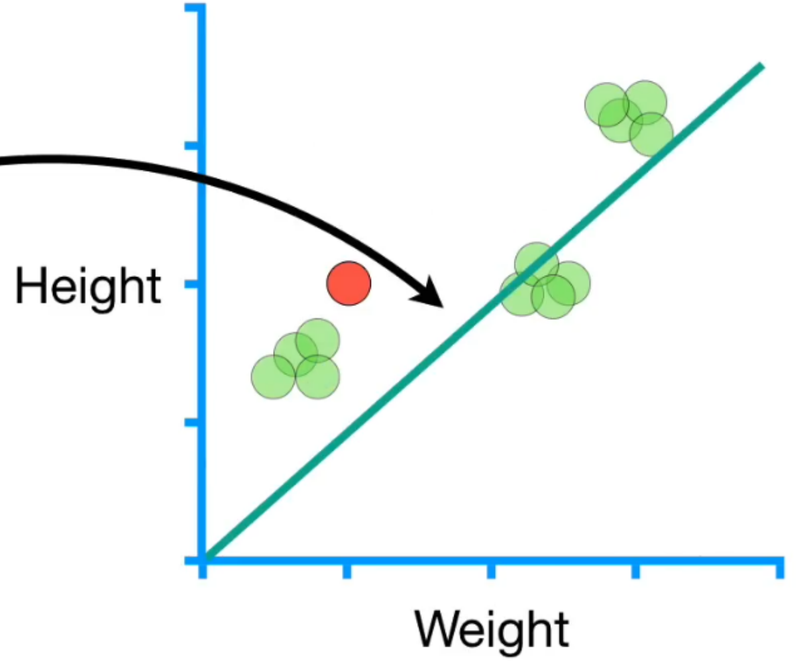
One cool thing about **Stochastic Gradient Descent** is that when we get new data…



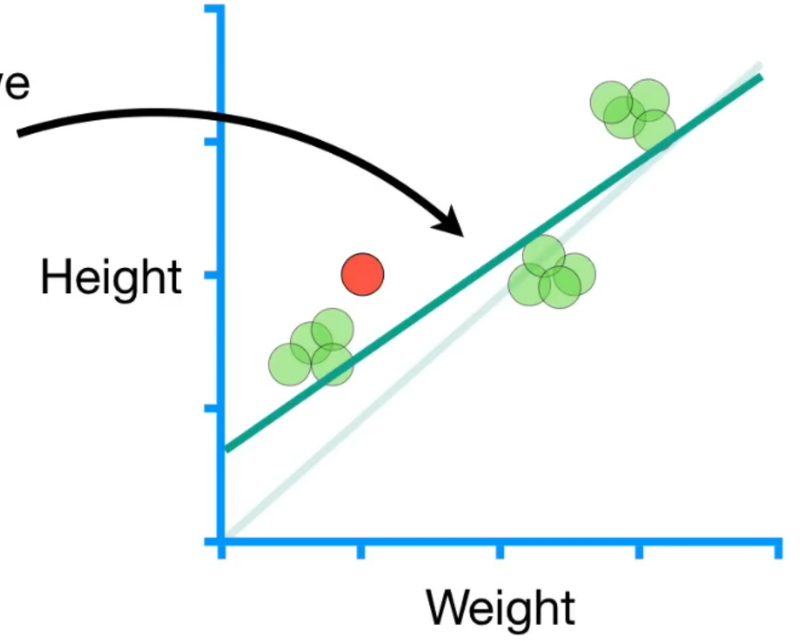
…we can easily use it to take another step for the parameter estimates without having to start from scratch.



In other words, we don't have to go all of the way back to the initial guesses for the **slope** and **intercept** and redo everything.



Instead, we pick up right where we left off and take one more step using the new sample.



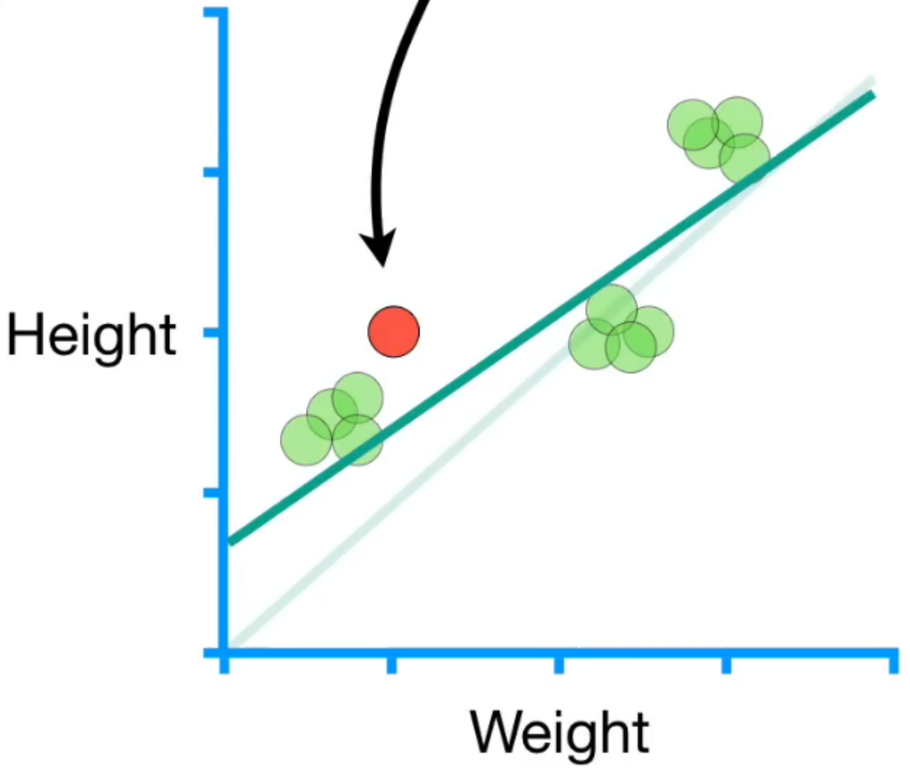
$$\frac{d}{d\,intercept}$$ Sum of squared residuals =

-2(**Height** - (0 + 1 × **Weight**)

So we plug in the **Weight** from the new sample, **1.1**…

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

-2 × **Weight**(**Height** - (0 + 1 × **Weight**))



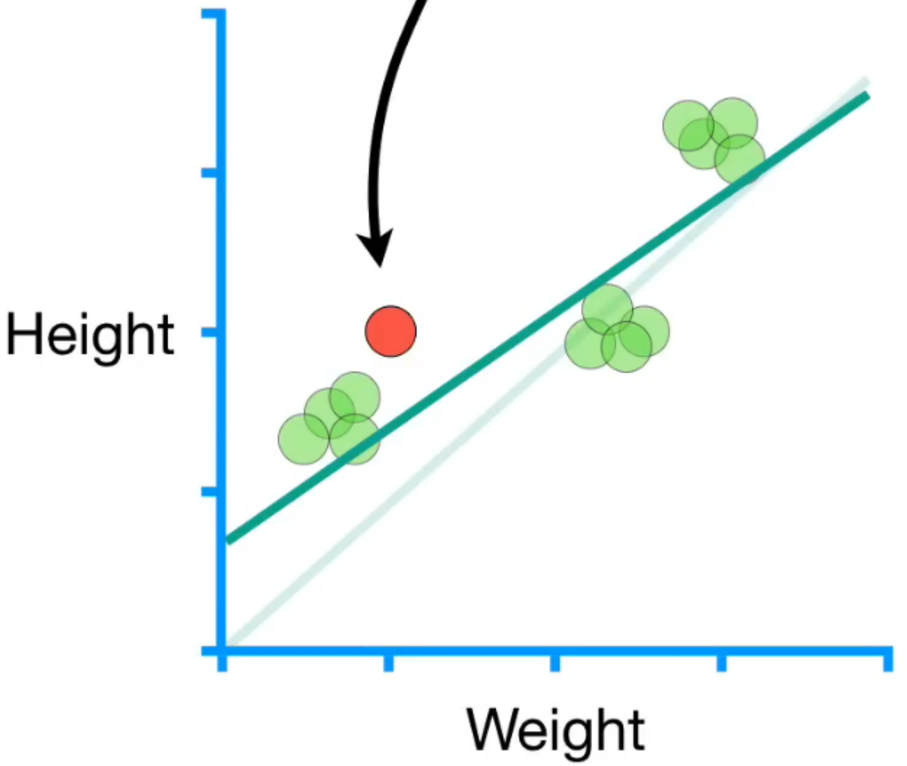
$\frac{d}{d\ intercept}$ Sum of squared residuals =

-2(**Height** - (0 + 1 × **1.1**)

…and the **Height**, **2**…

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

-2 × **1.1**(**Height** - (0 + 1 × **1.1**))



Plug in the slope and then multiply by the Learning Rate, 0.01

$$\frac{d}{d\,intercept}\text{Sum of squared residuals} = -2(\mathbf{2} - (0 + 1 \times \mathbf{1.1}) = -1.8$$

$$\textbf{Step Size}_{\text{Intercept}} = \textbf{Slope} \times \textbf{Learning Rate}$$

$$\textbf{Step Size}_{\text{Slope}} = \textbf{Slope} \times \textbf{Learning Rate}$$

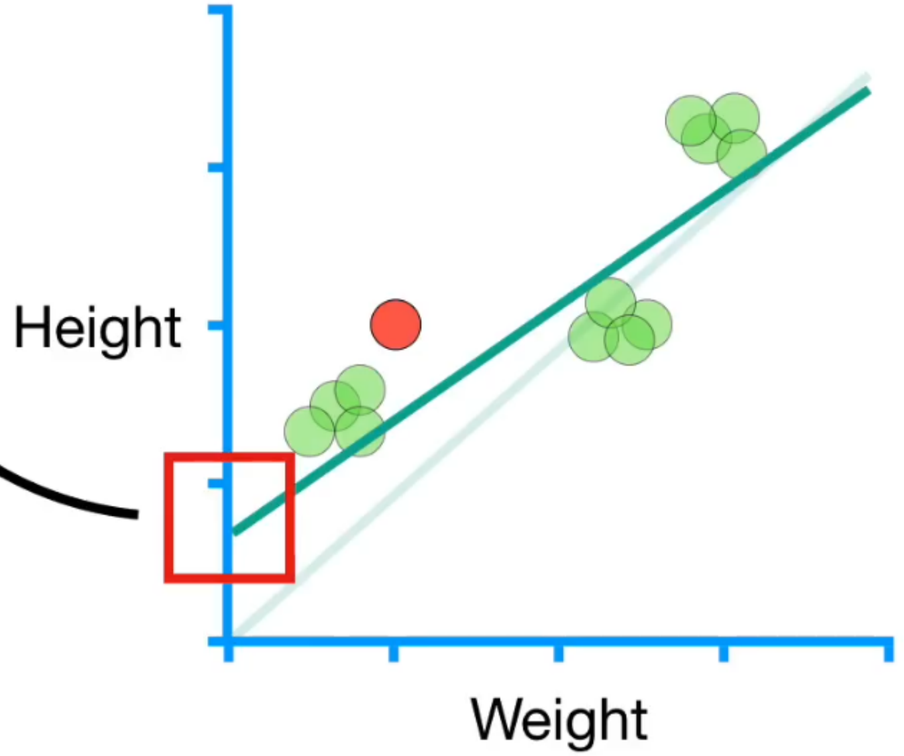
$$\frac{d}{d\,slope}\text{Sum of squared residuals} = -2 \times \mathbf{1.1}(\mathbf{2} - (0 + 1 \times \mathbf{1.1})) = -1.98$$

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

Step Size$_{\text{Intercept}}$ = -1.8 x 0.01 = **-0.018**

Step Size$_{\text{Slope}}$ = -1.98 x 0.01 = **-0.02**

…calculate the new **intercept,** not from the initial guess, but from the most recent estimate…



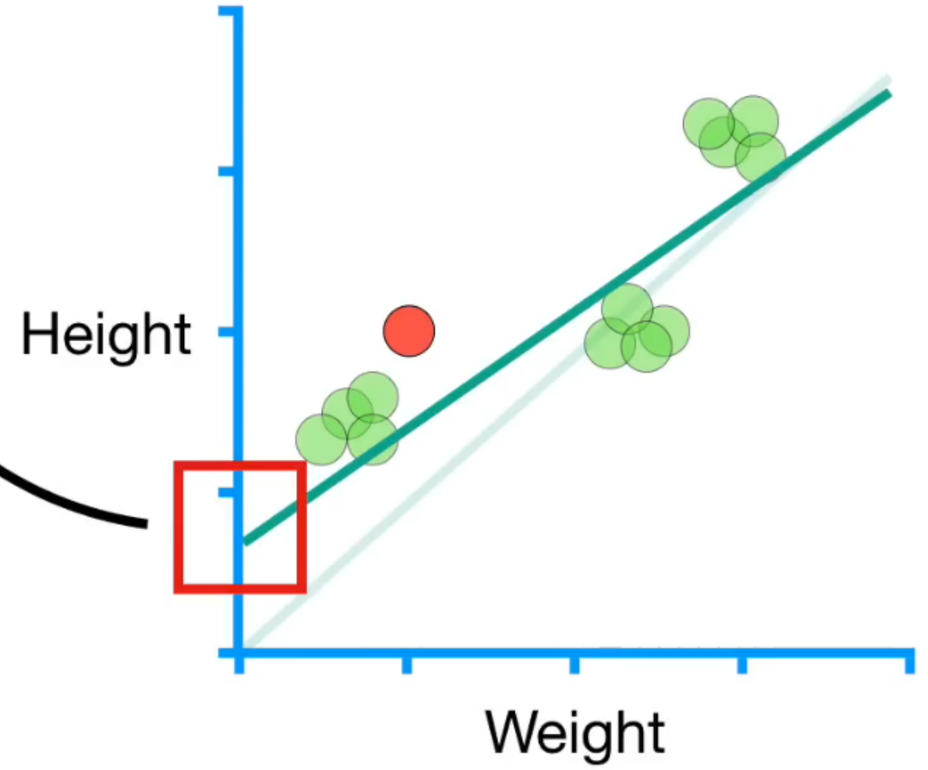
New Intercept = 0.86 - -0.018 = 0.878

New Intercept = 0.86 - **Step Size**

Step Size$_{\text{Intercept}}$ = -1.8 x 0.01 = **-0.018**

Step Size$_{\text{Slope}}$ = -1.98 x 0.01 = **-0.02**

…calculate the new **intercept,** not from the initial guess, but from the most recent estimate…



New Intercept = 0.86 - -0.018 = 0.878

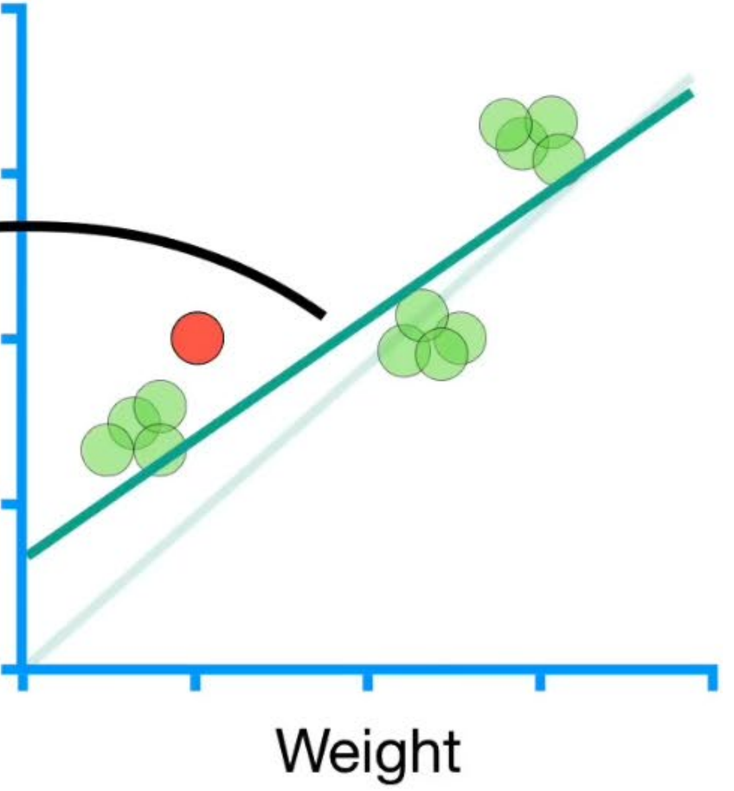
Step Size$_{\text{Intercept}}$ = -1.8 x 0.01 = **-0.018**

…and calculate the new **slope** from the most recent estimate…

Step Size$_{\text{Slope}}$ = -1.98 x 0.01 = **-0.02**

New Slope = 0.68 - **Step Size**

New Slope = 0.68 - -0.02 = 0.7



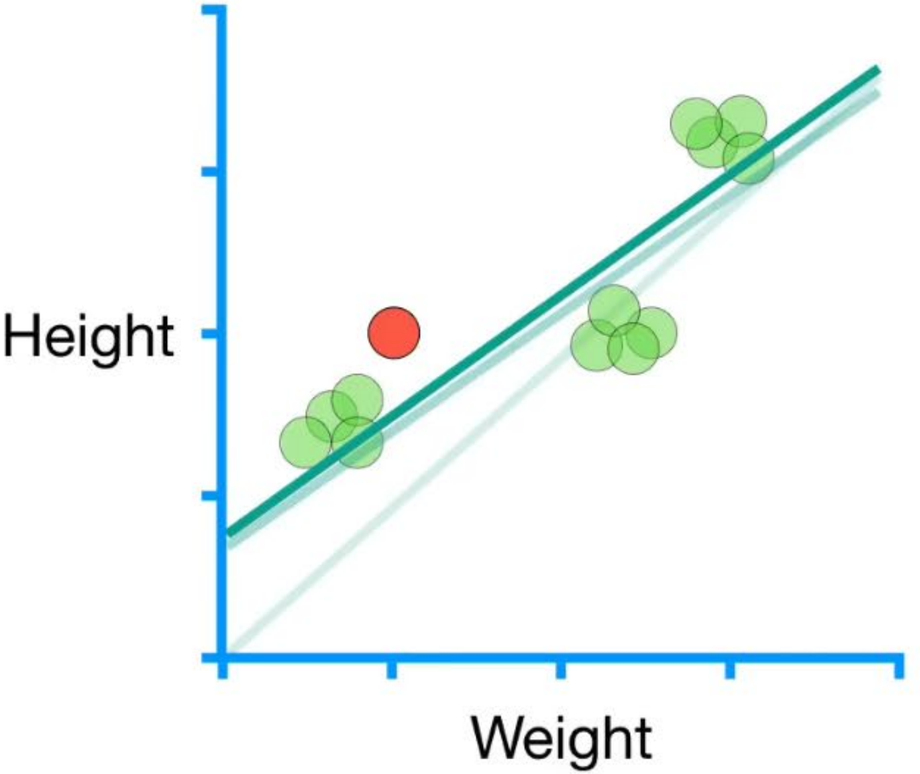
New Intercept = 0.86 - -0.018 = 0.878

Step Size$_{\text{Intercept}}$ = -1.8 x 0.01 = **-0.018**

Step Size$_{\text{Slope}}$ = -1.98 x 0.01 = **-0.02**

New Slope = 0.68 - -0.02 = 0.7

…and the new line has **intercept** = **0.878** and **slope** = **0.07**.



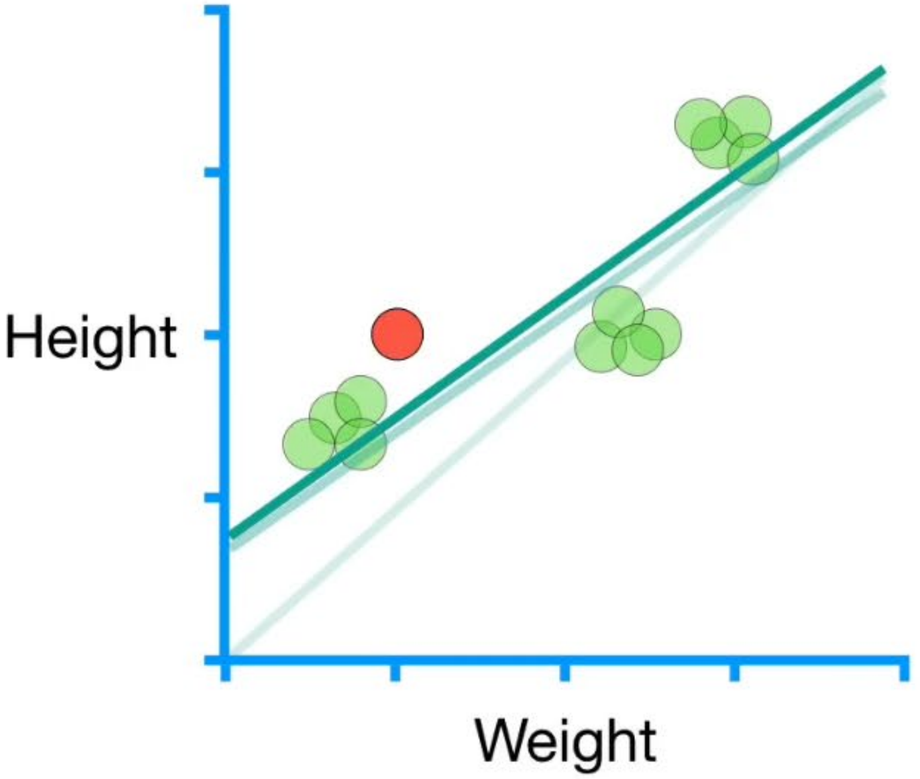
New Intercept = 0.86 - -0.018 = 0.878

$\text{Step Size}_{\text{Intercept}} = -1.8 \times 0.01 = \mathbf{-0.018}$

$\text{Step Size}_{\text{Slope}} = -1.98 \times 0.01 = \mathbf{-0.02}$

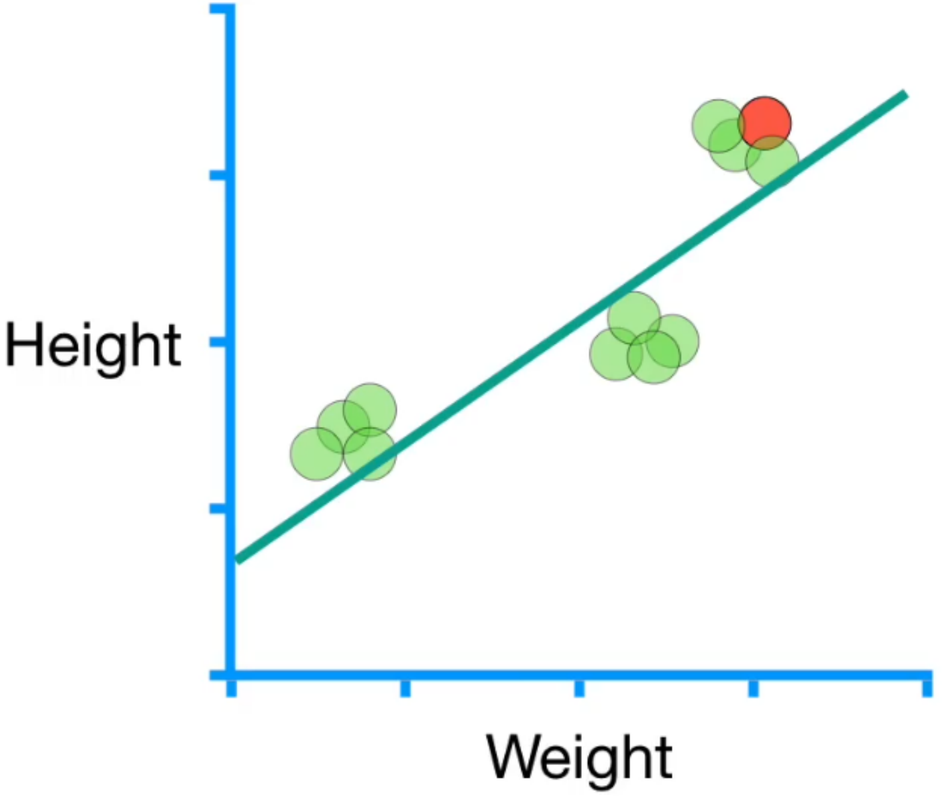
New Slope = 0.68 - -0.02 = 0.7

We updated the parameters for the line with just the new data.

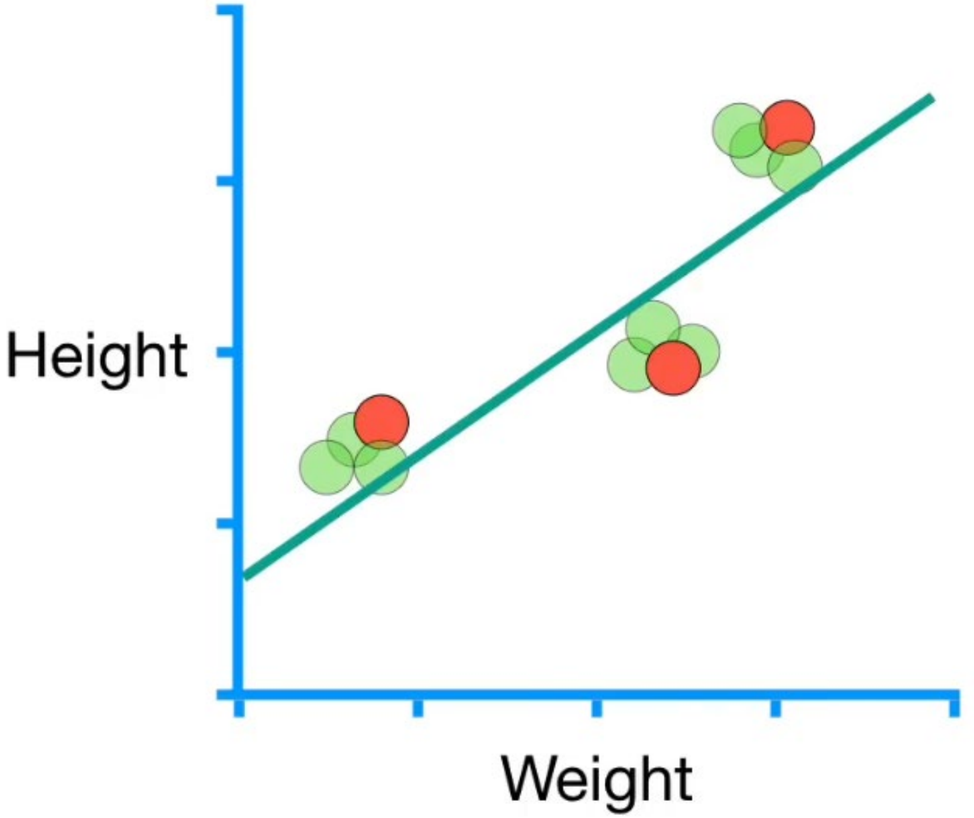


In Summary

Stochastic Gradient Descent is just like regular **Gradient Descent**, except it only looks at one sample per step…



…or a small subset, or
mini-batch, for each step.



$$\frac{d}{d\,gene1}\,\text{Loss Function()}$$

$$\frac{d}{d\,gene2}\,\text{Loss Function()}$$

$$\frac{d}{d\,gene3}\,\text{Loss Function()}$$

$$\frac{d}{d\,gene4}\,\text{Loss Function()}$$

$$\frac{d}{d\,gene5}\,\text{Loss Function()}$$

$$\frac{d}{d\,gene6}\,\text{Loss Function()}$$

$$\frac{d}{d\,gene7}\,\text{Loss Function()}$$

etc…etc…etc…

Stochastic Gradient Descent is great when we have tons of data and a lot of parameters.

In these situations, regular **Gradient Descent** may not be computationally feasible.

And it's cool that we can easily update the parameters when new data shows up.